

## PATENT ABSTRACTS OF JAPAN

(11)Publication number : 10-105347

(43)Date of publication of application : 24.04.1998

(51)Int.Cl.

G06F 3/06  
G06F 3/06

(21)Application number : 08-262147

(71)Applicant : HITACHI LTD

(22)Date of filing : 02.10.1996

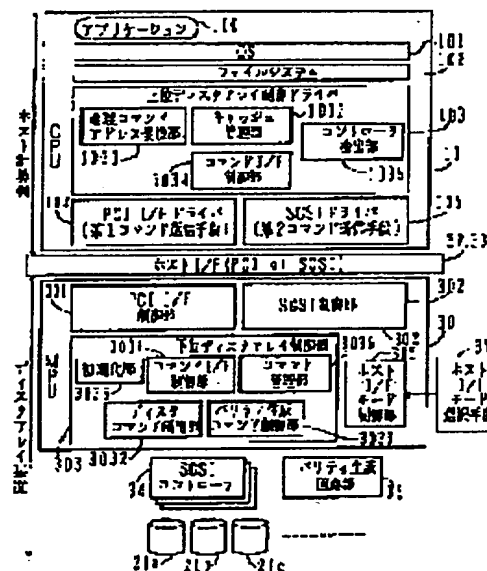
(72)Inventor : MATSUNAMI NAOTO  
OEDA TAKASHI  
KANEDA TAISUKE  
ARAKAWA TAKASHI  
YAGISAWA IKUYA  
TAKANO MASAHIRO

## (54) DISK ARRAY CONTROL SYSTEM

## (57)Abstract:

PROBLEM TO BE SOLVED: To provide a high-performance disk array control system whose performance is not limited even when an inexpensive micro processing unit(MPU) is used for a disk array controller.

SOLUTION: The CPU 10 of a host computer is equipped with a host disk array control driver 103 which generates a disk command from a logical command to a disk array device and also generates a parity generation command. The disk array controller is equipped with an MPU 30 which has lower throughput than a CPU that executes a control program in the disk array controller, a SCSI controller 34, and a parity generating circuit 35. The MPU 30 is equipped with a slave disk array control part 303 which receives the disk command and parity generation command issued by the host disk array control driver 103 and sends them to disks and the parity generating circuit 35.



## LEGAL STATUS

[Date of request for examination] 06.03.2000

[Date of sending the examiner's decision of rejection] 24.09.2003

[Kind of final disposal of application other than the examiner's decision of rejection or application converted registration]

[Date of final disposal for application]

[Patent number]

[Date of registration]

[Number of appeal against examiner's decision of rejection]

[Date of requesting appeal against examiner's decision of rejection]

[Date of extinction of right]

Copyright (C); 1998,2003 Japan Patent Office

**THIS PAGE BLANK (USPTO)**

(19) 日本国特許庁 (J P)

(12) 公開特許公報 (A)

(11) 特許出願公開番号

特開平10-105347

(43) 公開日 平成10年(1998) 4月24日

(51) Int.Cl.<sup>6</sup>

G 0 6 F 3/06

識別記号

5 4 0

3 0 5

F I

G 0 6 F 3/06

5 4 0

3 0 5 C

審査請求 未請求 請求項の数9 O L (全 38 頁)

(21) 出願番号 特願平8-262147

(22) 出願日 平成8年(1996)10月2日

(71) 出願人 000005108

株式会社日立製作所

東京都千代田区神田駿河台四丁目6番地

(72) 発明者 松並 直人

神奈川県川崎市麻生区王禅寺1099番地 株

式会社日立製作所システム開発研究所内

(72) 発明者 大枝 高

神奈川県川崎市麻生区王禅寺1099番地 株

式会社日立製作所システム開発研究所内

(72) 発明者 兼田 泰典

神奈川県川崎市麻生区王禅寺1099番地 株

式会社日立製作所システム開発研究所内

(74) 代理人 弁理士 春日 譲

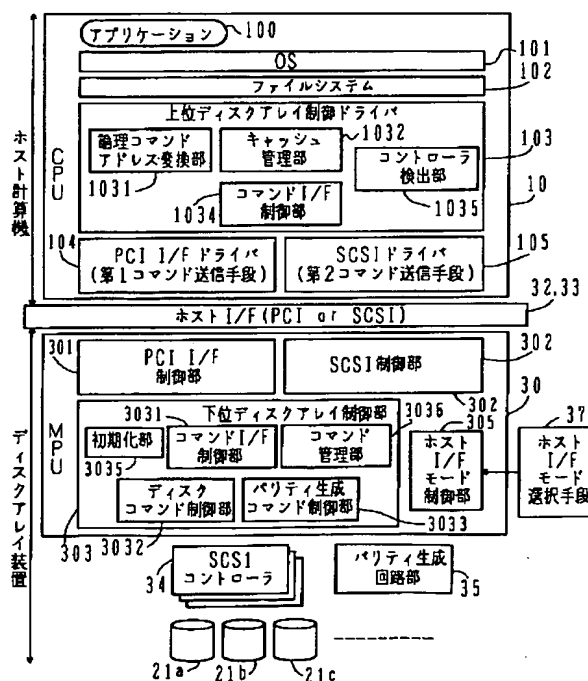
最終頁に続く

(54) 【発明の名称】 ディスクアレイ制御システム

(57) 【要約】 (修正有)

【課題】ディスクアレイコントローラに低価格なMPUを使用しても性能が制限されない高性能なディスクアレイ制御システムを提供する。

【解決手段】ホスト計算機のCPU10は、ディスクアレイ装置への論理コマンドからディスクコマンドを生成し、かつ、パリティ生成コマンドを生成する上位ディスクアレイ制御ドライバ103を備えている。ディスクアレイコントローラは、ディスクアレイコントローラの内部の制御プログラムを実行するCPUよりも処理能力の小さいMPU30と、SCSIコントローラ34と、パリティ生成回路35とを備える。MPU30は、上位ディスクアレイ制御ドライバ103が発行したディスクコマンド及びパリティ生成コマンドを受信してそれらをディスクやパリティ生成回路35に対して発行する下位ディスクアレイ制御部303を備えている。



(2)

特開平10-105347

**【特許請求の範囲】**

【請求項1】 プログラムを実行するCPUを有するホスト計算機と、

このホスト計算機にホストインターフェースを介して接続されるとともに、複数のディスクからなるディスクアレイと、このディスクアレイを制御するディスクアレイコントローラとを有するディスクアレイ装置から構成され、

上記CPUは、上記ディスクアレイ装置への論理コマンドから上記ディスクアレイを構成するディスクへのディスクコマンドを生成し、かつ、冗長データを生成するための冗長データ生成コマンドを生成する上位ディスクアレイ制御手段を備え、

上記ディスクアレイコントローラは、

上記ディスクアレイコントローラの内部の制御プログラムを実行する上記CPUよりも処理能力の小さいMPUと、

上記ディスクアレイを構成するディスクを接続するディスクインターフェースと、

上記ディスクアレイに格納する冗長データを生成する冗長データ生成手段とを備え、

上記MPUは、上記上位ディスクアレイ制御手段が発行した上記ディスクコマンド及び冗長データ生成コマンドを受信し、上記ディスクアレイを構成するディスクにディスクコマンドを発行し、また、上記冗長データ生成手段に冗長データ生成コマンドを発行する下位ディスクアレイ制御手段を備えることを特徴とするハイブリッドアレイ構成のディスクアレイ制御システム。

【請求項2】 プログラムを実行するCPUを有するホスト計算機と、

このホスト計算機にホストインターフェースを介して接続されるとともに、複数のディスクからなるディスクアレイと、このディスクアレイを制御するディスクアレイコントローラとを有するディスクアレイ装置から構成され、

上記ディスクアレイ装置の上記ホストインターフェースは、上記ホスト計算機の内部システムバスに直結するカードエッジを有する構造の第1ホストインターフェースと、

上記ホスト計算機とケーブルを介して接続する構造の第2ホストインターフェースとから構成され、

上記ディスクアレイ装置は、上記第1ホストインターフェースと上記第2ホストインターフェースのいずれか一方を選択するホストインターフェースモード選択手段を備え、内蔵型と外付け型を選択できることを特徴とするディスクアレイ制御システム。

【請求項3】 請求項2記載のディスクアレイ制御システムにおいて、

上記CPUは、さらに、

上記第1ホストインターフェースに上記CPUが生成し

たコマンドを送信し、上記ディスクアレイコントローラからの終了報告を受信する第1コマンド送信制御手段と、

上記第2ホストインターフェースに上記CPUが生成したコマンドを送信し、ディスクアレイコントローラからの終了報告を受信する第2コマンド送信制御手段と、上記第1若しくは第2コマンド送信制御手段を選択するコマンド送信選択手段を備え、

上記ディスクアレイコントローラは、さらに、

上記第1ホストインターフェースから上記CPUが生成したコマンドを受信し、その終了を上記ホスト計算機に報告する第1コマンド受信制御手段と、

上記第2ホストインターフェースから上記CPUが生成したコマンドを受信し、その終了をホストに報告する第2コマンド受信制御手段と、

上記ホストインターフェースモード選択手段により決定した上記第1若しくは第2ホストインターフェースに応じた上記第1若しくは第2コマンド受信制御手段を選択するコマンド受信選択手段とを備えたことを特徴とするディスクアレイ制御システム。

【請求項4】 請求項2記載のディスクアレイ制御システムにおいて、

上記ホストインターフェースモード選択手段が、上記第2ホストインターフェースを選択した際には、

上記第2ホストインターフェースと上記ホスト計算機はケーブルで接続し、

上記ホスト計算機の外部に設けたディスクアレイ筐体に上記ディスクアレイコントローラ及び上記ディスクアレイを搭載し、

上記ディスクアレイ筐体は、上記第1ホストインターフェースのカードエッジを接続する構造のコネクタと、電源と、クロック信号生成手段を備え、

上記ディスクアレイコントローラは、第1ホストインターフェースのカードエッジで上記コネクタに接続し、上記コネクタを介して上記電源から電力を供給し、上記クロック信号生成手段からクロック信号を供給することを特徴とするディスクアレイ制御システム。

【請求項5】 請求項4記載のディスクアレイ制御システムにおいて、

上記ディスクアレイ筐体は、上記ホストインターフェースモード選択手段と接続するコネクタを有し、

上記ホストインターフェースモード選択手段は、上記ディスクアレイ筐体のコネクタと接続した際には第2ホストインターフェース制御手段を選択することを特徴とするディスクアレイ制御システム。

【請求項6】 プログラムを実行するCPUを有するホスト計算機と、

このホスト計算機にホストインターフェースを介して接続されるとともに、複数のディスクからなるディスクアレイと、このディスクアレイを制御するディスクアレイ

(3)

特開平10-105347

コントローラとを有するディスクアレイ装置から構成され、  
上記CPU若しくは上記MPUは、上記ディスクアレイ装置への論理コマンドから上記ディスクアレイを構成するディスクへのディスクコマンドを生成し、かつ、冗長データを生成するための冗長データ生成コマンドを生成する上位ディスクアレイ制御手段を備え、  
上記ディスクアレイコントローラは、  
上記ディスクアレイコントローラの内部の制御プログラムを実行する上記CPUよりも処理能力の小さいMPUと、  
上記ディスクアレイを構成するディスクを接続するディスクインターフェースと、  
上記ディスクアレイに格納する冗長データを生成する冗長データ生成手段とを備え、  
上記MPUは、上記上位ディスクアレイ制御手段が発行した上記ディスクコマンド及び冗長データ生成コマンドを受信し、上記ディスクアレイを構成するディスクにディスクコマンドを発行し、また、上記冗長データ生成手段に冗長データ生成コマンドを発行する下位ディスクアレイ制御手段を備え、  
さらに、上記ディスクアレイコントローラは、上記CPU若しくは上記MPUが備える上記上位ディスクアレイ制御手段を選択するディスクアレイモード選択手段を備え、ハイブリッドアレイ構成とハードアレイ構成を選択できることを特徴とするディスクアレイ制御システム。  
【請求項7】 プログラムを実行するCPUを有する複数のホスト計算機と、  
これらのホスト計算機にホストインターフェースを介して接続されるとともに、複数のディスクからなるディスクアレイと、このディスクアレイを制御するディスクアレイコントローラとを有するディスクアレイ装置から構成され、  
上記複数のホスト計算機のそれぞれの上記CPUは、上記ディスクアレイ装置への論理コマンドから上記ディスクアレイを構成するディスクへのディスクコマンドを生成し、かつ、冗長データを生成するための冗長データ生成コマンドを生成する上位ディスクアレイ制御手段を備え、  
上記ディスクアレイコントローラは、  
上記ディスクアレイコントローラの内部の制御プログラムを実行する上記CPUよりも処理能力の小さいMPUと、  
上記ディスクアレイを構成するディスクを接続するディスクインターフェースと、  
上記ディスクアレイに格納する冗長データを生成する冗長データ生成手段とを備え、  
上記MPUは、上記上位ディスクアレイ制御手段が発行した上記ディスクコマンド及び冗長データ生成コマンドを受信し、上記ディスクアレイを構成するディスクにディスクコマンドを発行し、また、上記冗長データ生成手段に冗長データ生成コマンドを発行する下位ディスクアレイ制御手段を備え、  
上記ディスクアレイ装置は、第1の上記ホスト計算機に内蔵され、  
上記ディスクアレイ装置の上記ホストインターフェースは、上記第1のホスト計算機の内部システムバスに直結するカードエッジを有する構造の第1ホストインターフェースと、  
上記第1のホスト計算機以外の上記ホスト計算機とケーブルを介して接続する構造の第2ホストインターフェースとから構成されることを特徴とするディスクアレイ制御システム。

ディスクコマンドを発行し、また、上記冗長データ生成手段に冗長データ生成コマンドを発行する下位ディスクアレイ制御手段を備え、  
上記ディスクアレイ装置は、第1の上記ホスト計算機に内蔵され、  
上記ディスクアレイ装置の上記ホストインターフェースは、上記第1のホスト計算機の内部システムバスに直結するカードエッジを有する構造の第1ホストインターフェースと、  
上記第1のホスト計算機以外の上記ホスト計算機とケーブルを介して接続する構造の第2ホストインターフェースとから構成されることを特徴とするディスクアレイ制御システム。  
【請求項8】 プログラムを実行するCPUを有する第1のホスト計算機と、  
プログラムを実行するCPUを有する第2のホスト計算機と、  
上記第1及び第2のホスト計算機にホストインターフェースを介して接続されるとともに、複数のディスクからなるディスクアレイと、このディスクアレイを制御するディスクアレイコントローラとを有するディスクアレイ装置から構成され、  
上記第1の計算機の上記CPUは、上記ディスクアレイ装置への論理コマンドから上記ディスクアレイを構成するディスクへのディスクコマンドを生成し、かつ、冗長データを生成するための冗長データ生成コマンドを生成する上位ディスクアレイ制御手段を備え、  
上記第2の計算機の上記CPUは、上記ディスクアレイ装置への論理コマンドを上記ディスクアレイコントローラに発行するディスク制御手段を備え、  
上記ディスクアレイコントローラは、  
上記ディスクアレイコントローラの内部の制御プログラムを実行する上記CPUよりも処理能力の小さいMPUと、  
上記ディスクアレイを構成するディスクを接続するディスクインターフェースと、  
上記ディスクアレイに格納する冗長データを生成する冗長データ生成手段とを備え、  
上記MPUは、上記上位ディスクアレイ制御手段が発行した上記ディスクコマンド及び冗長データ生成コマンドを受信し、上記ディスクアレイを構成するディスクにディスクコマンドを発行し、また、上記冗長データ生成手段に冗長データ生成コマンドを発行する下位ディスクアレイ制御手段を備え、  
上記ディスクアレイ装置は、第1の上記ホスト計算機に内蔵され、  
上記ディスクアレイ装置の上記ホストインターフェースは、上記第1のホスト計算機の内部システムバスに直結するカードエッジを有する構造の第1ホストインターフェースと、

(4)

特開平10-105347

上記第2のホスト計算機とケーブルを介して接続する構造の第2ホストインターフェースとから構成されることを特徴とするディスクアレイ制御システム。

【請求項9】 ホスト計算機にホストインターフェースを介して接続されるとともに、複数のディスクからなるディスクアレイを制御するディスクアレイコントローラにおいて、

上記ディスクアレイコントローラの内部の制御プログラムを実行するMPUと、

上記ディスクアレイを構成するディスクを接続するディスクインターフェースと、

上記ディスクアレイに格納する冗長データを生成する冗長データ生成手段とを備え、

上記ホストインターフェースは、上記ホスト計算機の内部システムバスに直結するカードエッジを有する構造の第1ホストインターフェースと、

上記ホスト計算機とケーブルを介して接続する構造の第2ホストインターフェースとから構成されることを特徴とするディスクアレイコントローラ。

【発明の詳細な説明】

【0001】

【発明の属する技術分野】本発明は、ディスクアレイを制御するディスクアレイ制御システムに関する。

【0002】

【従来の技術】ディスクアレイは、複数のディスクを並列に動作させることで、単体のディスクに比べ高速化を実現する技術である。しかし、ディスクをn台並べた場合、その故障確率は、n倍に悪化する。

【0003】そこで、高速化と高信頼性を両立するための技術として、「RAID (Redundant Arrays of Inexpensive Disks)」が知られている。RAIDは、例えば、"A Case for Redundant Arrays of Inexpensive Disks (RAID)"; In Proc. ACM SIGMOD, June 1988 (カリフォルニア大学バークレー校発行)に記載されている。

【0004】RAIDは、複数のディスクを並列に動作させることで高速制御を実現し、また、パリティと呼ぶ冗長データをパリティディスクと呼ぶ特定のディスクに格納することにより、万一、データを格納する1台のディスクが故障しても、他のディスクとパリティディスクのパリティとから故障したディスクのデータを再現することができ、耐ディスク障害信頼性を高めることができるディスク制御の方法である。

【0005】RAIDは、そのパリティの格納の方法によりレベル1から5がある。レベル4のRAID型ディスクアレイでは、例えば、ディスクが5台あり、4台がデータディスク0～3、1台がパリティディスクとすると、データをディスク0、ディスク1、ディスク2、デ

ィスク3の順番で、ある一定のデータブロック毎に順に分散して格納する。このデータブロック単位のことをストライプと称し、この分散する制御のことをストライピングと称する。

【0006】ディスク0～3に格納した同一ストライプのデータD0～3の排他的論理和（以下XORと称する）を計算することで、

$P = D0 + D1 + D2 + D3$  （ただし、+はXOR演算を示す）

パリティPを生成する事ができる。また、D0を格納するディスク0が故障した際には、

$D0 = D1 + D2 + D3 + P$

により、故障したディスク0のD0を再現できる。

【0007】また、RAIDをはじめとするディスクアレイの制御方法には、「ハードアレイ」方式、「ソフトウェア」方式と言われる方法がある。「ハードアレイ」方式は、ディスクアレイ制御専用のコントローラを備え、そのコントローラ上のCPUで、上記のようなディスクアレイのデータ分散制御を行う方式である。「ハードアレイ」は、例えば、米国特許第5,249,279号明細書に記載されている。ここで、コントローラ上のCPUを、ホスト計算機のCPUと区別するため、MPU (Micro Processing Unit) と称することにする。ただし、MPUは、1つで構成しても複数で構成してもよいので、これらを分け隔て無く同一の呼称を用いることにする。

【0008】「ハードアレイ」方式では、ホスト計算機は、ディスクアレイを1台の大容量の仮想ディスクとしてとらえ、仮想ディスクに対してリード/ライト等のコマンドを発行する。ディスクアレイコントローラはこのコマンドを受信し、ディスクアレイを構成する各ディスクへのコマンドに変換する。この方式は専用のハード（コントローラ）を有することから「ハードアレイ」と称されている。

【0009】一方、「ソフトウェア」方式は、ディスクアレイ専用のコントローラを持たず、ホスト計算機に複数台のディスクを接続し、CPUでディスクアレイのデータ分散制御を行うものである。ホスト計算機のCPUのソフトウェアでディスクアレイ制御をおこなうことから「ソフトウェア」と称されている。

【0010】

【発明が解決しようとする課題】最初に従来の第1の問題点について説明する。従来の「ハードアレイ」方式のディスクアレイでは、ロジカルアクセスリクエストをディスクのコマンドへの変換処理からディスクのコマンド処理まで、すべてのディスクアレイ制御を、ディスクアレイコントローラのMPUで行う方式である。この方法では、多数のディスクを接続し、多数のロジカルアクセスリクエストを処理しようとする、MPUの能力でその処理リクエスト数が限定されてしまうという問題があ

(5)

特開平10-105347

る。これは、一般に、ディスクアレイコントローラに搭載するMPUの能力は、ホスト計算機のCPU能力に比べ非常に低いためである。

【0011】また、従来の「ソフトアレイ」方式のディスクアレイでは、処理能力の高いCPUで全てのディスクアレイ処理を行うので、専用のMPUをもつ「ハードアレイ」より高性能が期待できる。しかし、CPUでは、負荷の重いアプリケーションや、ネットワーク、グラフィック等の処理を並行して実行しなくてはならないので、負荷の重いディスクアレイ処理をCPUで実行することにより、CPU負荷率が上昇し、本来のアプリケーション処理を妨げてしまうという問題がある。さらに、「ソフトアレイ」はホスト計算機のOSのファイルシステムや、デバイスドライバで実現するが、このため、ホスト計算機の種類やOSの種類に依存してしまうという問題がある。

【0012】次に、従来の第2の問題点について説明する。また、従来の「ハードアレイ」方式は、ディスクアレイコントローラをホスト計算機に内蔵する方式とホスト計算機に外付けする方式がある。内蔵方式は、ホストI/Fに高速なホストバスを用い、また、SCSI等のオーバヘッドが無いので高性能であり、さらに、ホスト計算機にディスクアレイを内蔵できるので小型化に適するが、その反面、ホストバスの種類に依存し、さらにロジカルディスクアクセスリクエストを発行する手段が、ホスト計算機のOSに依存してしまい、接続性が低いという問題がある。

【0013】一方、外付け方式は、ホストI/Fに、SCSI等のディスク接続バスを用いるので、ホスト計算機のホストバスやOSに依存しないため、接続性が高いが、その反面、SCSIのオーバヘッドが大きく、転送性能も低いので性能が劣るという問題がある。このように両方式は一長一短の面があり、ホスト計算機と要求性能によりいずれかを選択する必要がある。このため、それぞれのホストI/Fを有するディスクアレイコントローラを別々に製作する必要がある、開発コストが高くなるという問題がある。

【0014】次に、従来の第3の問題点について説明する。内蔵方式のハードアレイにおいては、ディスクアレイコントローラをホスト計算機に内蔵するので、複数台のホスト計算機が1台のディスクアレイを共用する構成を実現できないという問題がある。すなわち、複数台のホスト計算機で分散処理をすることで処理の高速化を図る「クラスタ構成」や、1台のホスト計算機が故障した際に待機するホスト計算機に処理を引き継ぐ「スタンバイ構成」を実現できないという問題がある。

【0015】本発明の第1の目的は、上記第1の問題点を解決して、ディスクアレイコントローラに低価格なMPUを使用しても、MPUの能力で性能が制限されない高性能なディスクアレイ制御システムを提供することに

ある。

【0016】本発明の第2の目的は、上記第2の問題点を解決して、ホストバスに接続してホスト計算機に内蔵する「内蔵型ディスクアレイ」と、SCSI等のディスク接続バスでホスト計算機に外付け接続する「外付け型ディスクアレイ」を、唯一つのディスクアレイコントローラで実現することができるディスクアレイ制御システムを提供することにある。

【0017】さらに、本発明の第3の目的は、ホスト計算機のホストバスやOSや、動作させるアプリケーションの要求性能に応じて、「ハードアレイ」方式のディスクアレイと、本発明の第1の目的を達成する高性能なディスクアレイを、選択的に切り換えることができるディスクアレイ制御システムを提供することにある。

【0018】本発明の第4の目的は、上記第3の問題点を解決して、内蔵型ディスクアレイにおいても、クラスタ構成やスタンバイ構成等の複数台のホスト計算機で1台のディスクアレイを共用できるディスクアレイ制御システムを提供することにある。

【0019】本発明の第5の目的は、ホスト計算機にディスクアレイを内蔵したときに、そのホスト計算機自体を一つのディスクアレイ装置として構成し得るディスクアレイ制御システムを提供することにある。

【0020】

【課題を解決するための手段】上記第1の目的を実現するために、本発明は、プログラムを実行するCPUを有するホスト計算機と、このホスト計算機にホストインターフェースを介して接続されるとともに、複数のディスクからなるディスクアレイと、このディスクアレイを制御するディスクアレイコントローラとを有するディスクアレイ装置から構成され、上記CPUは、上記ディスクアレイ装置への論理コマンドから上記ディスクアレイを構成するディスクへのディスクコマンドを生成し、かつ、冗長データを生成するための冗長データ生成コマンドを生成する上位ディスクアレイ制御手段を備え、上記ディスクアレイコントローラは、上記ディスクアレイコントローラの内部の制御プログラムを実行する上記CPUよりも処理能力の小さいMPUと、上記ディスクアレイを構成するディスクを接続するディスクインターフェースと、上記ディスクアレイに格納する冗長データを生成する冗長データ生成手段とを備え、上記MPUは、上記上位ディスクアレイ制御手段が発行した上記ディスクコマンド及び冗長データ生成コマンドを受信し、上記ディスクアレイを構成するディスクにディスクコマンドを発行し、また、上記冗長データ生成手段に冗長データ生成コマンドを発行する下位ディスクアレイ制御手段を備えるようにしたものであり、かかる構成により、高性能化が図り得るものとなる。

【0021】上記第2の目的を達成するために、本発明は、プログラムを実行するCPUを有するホスト計算機

(6)

特開平10-105347

と、このホスト計算機にホストインターフェースを介して接続されるとともに、複数のディスクからなるディスクアレイと、このディスクアレイを制御するディスクアレイコントローラとを有するディスクアレイ装置から構成され、上記ディスクアレイ装置のホストインターフェースは、上記ホスト計算機の内部システムバスに直結するカードエッジを有する構造の第1ホストインターフェースと、上記ホスト計算機とケーブルを介して接続する構造の第2ホストインターフェースとから構成され、上記ディスクアレイ装置は、上記第1ホストインターフェースと上記第2ホストインターフェースのいずれか一方を選択するホストインターフェースモード選択手段を備えるようにしたものであり、かかる構成により、内蔵型と外付け型を選択し得るものとなる。

【0022】また、上記ディスクアレイ制御システムにおいて、好ましくは、上記CPUは、さらに、上記第1ホストインターフェースに上記CPUが生成したコマンドを送信し、上記ディスクアレイコントローラからの終了報告を受信する第1コマンド送信制御手段と、上記第2ホストインターフェースに上記CPUが生成したコマンドを送信し、ディスクアレイコントローラからの終了報告を受信する第2コマンド送信制御手段と、上記第1若しくは第2コマンド送信制御手段を選択するコマンド送信選択手段を備え、上記ディスクアレイコントローラは、さらに、上記第1ホストインターフェースから上記CPUが生成したコマンドを受信し、その終了を上記ホスト計算機に報告する第1コマンド受信制御手段と、上記第2ホストインターフェースから上記CPUが生成したコマンドを受信し、その終了をホストに報告する第2コマンド受信制御手段と、上記ホストインターフェースモード選択手段により決定した上記第1若しくは第2ホストインターフェースに応じた上記第1若しくは第2コマンド受信制御手段を選択するコマンド受信選択手段とを備えるようにしたものである。

【0023】上記ディスクアレイ制御システムにおいて、好ましくは、上記ホストインターフェースモード選択手段が、上記第2ホストインターフェースを選択した際には、上記第2ホストインターフェースと上記ホスト計算機はケーブルで接続し、上記ホスト計算機の外部に設けたディスクアレイ筐体に上記ディスクアレイコントローラ及び上記ディスクアレイを搭載し、上記ディスクアレイ筐体は、上記第1ホストインターフェースのカードエッジを接続する構造のコネクタと、電源と、クロック信号生成手段を備え、上記ディスクアレイコントローラは、第1ホストインターフェースのカードエッジで上記コネクタに接続し、上記コネクタを介して上記電源から電力を供給し、上記クロック信号生成手段からクロック信号を供給するようにしたものである。

【0024】上記ディスクアレイ制御システムにおいて、好ましくは、上記ディスクアレイ筐体は、上記ホス

トインターフェースモード選択手段と接続するコネクタを有し、上記ホストインターフェースモード選択手段は、上記ディスクアレイ筐体のコネクタと接続した際には第2ホストインターフェース制御手段を選択するようにしたものである。

【0025】上記第3の目的を達成するために、本発明は、プログラムを実行するCPUを有するホスト計算機と、このホスト計算機にホストインターフェースを介して接続されるとともに、複数のディスクからなるディスクアレイと、このディスクアレイを制御するディスクアレイコントローラとを有するディスクアレイ装置から構成され、上記CPU若しくは上記MPUは、上記ディスクアレイ装置への論理コマンドから上記ディスクアレイを構成するディスクへのディスクコマンドを生成し、かつ、冗長データを生成するための冗長データ生成コマンドを生成する上位ディスクアレイ制御手段を備え、上記ディスクアレイコントローラは、上記ディスクアレイコントローラの内部の制御プログラムを実行する上記CPUよりも処理能力の小さいMPUと、上記ディスクアレイを構成するディスクを接続するディスクインターフェースと、上記ディスクアレイに格納する冗長データを生成する冗長データ生成手段とを備え、上記MPUは、上記上位ディスクアレイ制御手段が発行した上記ディスクコマンド及び冗長データ生成コマンドを受信し、上記ディスクアレイを構成するディスクにディスクコマンドを発行し、また、上記冗長データ生成手段に冗長データ生成コマンドを発行する下位ディスクアレイ制御手段を備え、さらに、上記ディスクアレイコントローラは、上記CPU若しくは上記MPUが備える上記上位ディスクアレイ制御手段を選択するディスクアレイモード選択手段を備えるようにしたものであり、かかる構成により、ハイブリッドアレイ構成とハードアレイ構成を選択し得るものとなる。

【0026】上記第4の目的を達成するために、本発明は、プログラムを実行するCPUを有する複数のホスト計算機と、これらのホスト計算機にホストインターフェースを介して接続されるとともに、複数のディスクからなるディスクアレイと、このディスクアレイを制御するディスクアレイコントローラとを有するディスクアレイ装置から構成され、上記複数のホスト計算機のそれぞれの上記CPUは、上記ディスクアレイ装置への論理コマンドから上記ディスクアレイを構成するディスクへのディスクコマンドを生成し、かつ、冗長データを生成するための冗長データ生成コマンドを生成する上位ディスクアレイ制御手段を備え、上記ディスクアレイコントローラは、上記ディスクアレイコントローラの内部の制御プログラムを実行する上記CPUよりも処理能力の小さいMPUと、上記ディスクアレイを構成するディスクを接続するディスクインターフェースと、上記ディスクアレイに格納する冗長データを生成する冗長データ生成手段



(7)

特開平10-105347

とを備え、上記MPUは、上記上位ディスクアレ制御手段が発行した上記ディスクコマンド及び冗長データ生成コマンドを受信し、上記ディスクアレを構成するディスクにディスクコマンドを発行し、また、上記冗長データ生成手段に冗長データ生成コマンドを発行する下位ディスクアレ制御手段を備え、上記ディスクアレ装置は、第1の上記ホスト計算機に内蔵され、上記ディスクアレ装置の上記ホストインターフェースは、上記第1のホスト計算機の内部システムバスに直結するカードエッジを有する構造の第1ホストインターフェースと、上記第1のホスト計算機以外の上記ホスト計算機とケーブルを介して接続する構造の第2ホストインターフェースとから構成するようにしたものである。

【0027】上記第5の目的を達成するために、本発明は、プログラムを実行するCPUを有する第1のホスト計算機と、プログラムを実行するCPUを有する第2のホスト計算機と、上記第1及び第2のホスト計算機にホストインターフェースを介して接続されるとともに、複数のディスクにからなるディスクアレと、このディスクアレを制御するディスクアレコントローラとを有するディスクアレ装置から構成され、上記第1の計算機の上記CPUは、上記ディスクアレ装置への論理コマンドから上記ディスクアレを構成するディスクへのディスクコマンドを生成し、かつ、冗長データを生成するための冗長データ生成コマンドを生成する上位ディスクアレ制御手段を備え、上記第2の計算機の上記CPUは、上記ディスクアレ装置への論理コマンドを上記ディスクアレコントローラに発行するディスク制御手段を備え、上記ディスクアレコントローラは、上記ディスクアレコントローラの内部の制御プログラムを実行する上記CPUよりも処理能力の小さいMPUと、上記ディスクアレを構成するディスクを接続するディスクインターフェースと、上記ディスクアレに格納する冗長データを生成する冗長データ生成手段とを備え、上記MPUは、上記上位ディスクアレ制御手段が発行した上記ディスクコマンド及び冗長データ生成コマンドを受信し、上記ディスクアレを構成するディスクにディスクコマンドを発行し、また、上記冗長データ生成手段に冗長データ生成コマンドを発行する下位ディスクアレ制御手段を備え、上記ディスクアレ装置は、第1の上記ホスト計算機に内蔵され、上記ディスクアレ装置の上記ホストインターフェースは、上記第1のホスト計算機の内部システムバスに直結するカードエッジを有する構造の第1ホストインターフェースと、上記第2のホスト計算機とケーブルを介して接続する構造の第2ホストインターフェースとから構成するようにしたものである。

【0028】また、本発明は、ホスト計算機にホストインターフェースを介して接続されるとともに、複数のディスクにからなるディスクアレを制御するディスクアレ

コントローラにおいて、上記ディスクアレコントローラの内部の制御プログラムを実行するMPUと、上記ディスクアレを構成するディスクを接続するディスクインターフェースと、上記ディスクアレに格納する冗長データを生成する冗長データ生成手段とを備え、上記ホストインターフェースは、上記ホスト計算機の内部システムバスに直結するカードエッジを有する構造の第1ホストインターフェースと、上記ホスト計算機とケーブルを介して接続する構造の第2ホストインターフェースとから構成するようにしたものである。

【0029】

【発明の実施の形態】以下、図1～図19を用いて、本発明の一実施形態によるディスクアレ制御システムについて説明する。

【0030】[構成の説明]最初に、図1を用いて、本発明の一実施形態による内蔵型ハイブリッドアレ構成のディスクアレ制御システムの全体構成について説明する。図1は、本発明の一実施形態によるディスクアレ制御システムのハードウェアの全体構成のブロック図である。

【0031】ホスト計算機1は、CPU10と、主記憶メモリ11と、システム制御手段12と、PCIバス13と、PCIコネクタ14と、ディスクアレ装置2によって構成されている。

【0032】システム制御手段12は、CPU10とPCIバス13と主記憶メモリ11との間のデータ転送を制御する。PCIバス13は、ホストバス若しくはシステムバスである。PCIコネクタ14は、PCIバス13にPCI拡張ボードを接続する。ディスクアレ装置2は、PCIコネクタ14を介して、PCIバス13に接続されている。ディスクアレ装置2は、ホスト計算機1の中に内蔵されており、この構成を、「内蔵型」と称する。

【0033】ディスクアレ装置2は、少なくとも1台以上のディスク21a、21b、21c、21d、21e、21f、…と、ディスクアレコントローラ3によって構成されている。ディスクアレコントローラ3は、複数台のディスク21a、21b、…を制御して、これらのディスクをディスクアレとして機能させるディスクアレコントローラ3は、MPU30と、メモリ42と、PCI I/Fコントローラ321と、SCSI33、34a、34bと、パリティ生成回路35と、キャッシュメモリ36と、ホストI/Fモード選択手段37と、PCIバス38とから構成されている。

【0034】MPU30は、ディスクアレコントローラ3の制御プログラムを実行するものである。メモリ42は、MPU30が実行するプログラムやデータを格納している。

【0035】PCI I/Fコントローラ321は、ホスト計算機1の第1ホストI/FであるPCI I/F3

(8)

特開平10-105347

2を制御する。PCI I/F 32は、PCIコネクタ14とPCI I/Fコントローラ321によって構成される。また、PCI I/Fコントローラ321は、DMAコントローラ（図示せず）を内蔵している。

【0036】SCSI 33は、ホスト計算機1と接続するための第2ホストI/Fである。SCSI 33は、PCIカードエッジ332と、SCSIコントローラ331によって構成されている。SCSIコントローラ331は、SCSI 33を制御する。SCSIコントローラ331は、DMAコントローラ（図示せず）を内蔵している。PCIカードエッジ332は、SCSIケーブルを接続するSCSIコネクタであるが、本実施形態においては、SCSIケーブルは接続されていない。

【0037】SCSI 34aは、少なくとも1台以上のディスク21a, 21b, 21c, ...を接続するためのディスクI/Fである。SCSI 34aは、SCSIコントローラ34a1と、SCSIコネクタ40a1によって構成されている。SCSIコントローラ34a1は、SCSI 34aを制御する。SCSIコントローラ34a1は、DMAコントローラ（図示せず）を内蔵している。SCSIコネクタ40a1は、SCSIケーブルを接続するコネクタである。

【0038】SCSI 34bは、少なくとも1台以上のディスク21d, 21e, 21f, ...を接続するためのディスクI/Fである。SCSI 34bは、SCSIコントローラ34b1と、SCSIコネクタ40b1によって構成されている。SCSIコントローラ34b1は、SCSI 34bを制御する。SCSIコントローラ34b1は、DMAコントローラ（図示せず）を内蔵している。SCSIコネクタ40b1は、SCSIケーブルを接続するコネクタである。

【0039】パリティ生成回路35は、ディスクアレイに格納する冗長データであるパリティデータを生成する。キャッシュメモリ36は、ディスク21a, 21b, 21c, 21d, 21e, 21f, ...のデータを一時的に保持する。

【0040】ホストI/Fモード選択手段37は、ホスト計算機1とディスクアレイコントローラ3の接続を上記第1ホストI/FであるPCI I/F 32を用いるか、第2ホストI/FであるSCSI 33を用いるかを選択するものである。

【0041】PCIバス38は、ディスクアレイコントローラ3の内部のバスである。

【0042】同図において、ホスト計算機1のホストバス及びディスクアレイコントローラ3の内部バスにPCIバス13, 38を用いているが、他のバスを用いてもよい。

【0043】また、ディスクI/F 34a, 34bにSCSIを用いているがこれもこの限りではない。また、SCSIコントローラ34の数を2としているが、これ

はそれ以上でも以下でもよい。また、PCI I/Fコントローラ321とMPU30は別の構成にしているが、MPU30がPCI I/Fを備えているものであってもよい。また、キャッシュメモリ36の制御回路は、キャッシュメモリ36に含まれているものとし、パリティ生成回路35は独立した構成にしているが、パリティ生成回路をキャッシュメモリ制御回路と一体化して、キャッシュメモリ36内に含まれるものとしてもよい。このように、本実施形態は、様々な構成のバリエーションに適用できるものであり、本実施形態は本発明の範囲を限定するものではない。

【0044】次に、図2を用いて、本発明の一実施形態による内蔵型ハイブリッドアレイ構成のディスクアレイ制御システムのCPU10及びMPU30で実行するプログラムの構成について説明する。図2は、本発明の一実施形態によるディスクアレイ制御システムのソフトウェアの全体構成のブロック図である。

【0045】本実施形態においては、ディスクアレイ処理を、ホスト計算機1のCPU10とディスクコントローラ3のMPU30とで分散しており、この処理構成を「ハイブリッドアレイ」構成と称する。

【0046】CPU10が実行するプログラムは、アプリケーション100と、OS101と、ファイルシステム102と、上位ディスクアレイ制御ドライバ103と、PCI I/Fドライバ104と、SCSIドライバ105から構成されている。

【0047】ここで、上位ディスクアレイ制御ドライバ103は、論理コマンドアドレス変換部1031と、キャッシュ管理部1032と、コマンドI/F制御部1034と、コントローラ検出部1035とから構成されている。論理コマンドアドレス変換部1031は、ファイルシステム102から受信したディスクアレイへの論理コマンドをディスクアレイ装置2を構成する少なくとも1台以上のディスク21へのコマンドに変換するものである。キャッシュ管理部1032は、キャッシュメモリ36の管理を行うものである。コマンドI/F制御部1034は、PCI I/Fドライバ104およびSCSIドライバ105のいずれか選択されたドライバとの間でコマンド送信等のやりとりを行うものである。コントローラ検出部1035は、PCI I/F 32とSCSI 33のどちらのホストI/Fを使用するかを選択し、ディスクアレイコントローラの検出、初期化を行うものである。

【0048】PCI I/Fドライバ104は、PCI I/F 32を用いてディスクアレイコントローラ3にコマンドを送信するためのものであり、第1コマンド送信手段である。SCSIドライバ105は、SCSI 33を用いてディスクアレイコントローラ3にコマンドを送信するためのものであり、第2コマンド送信手段である。

(9)

特開平10-105347

【0049】また、MPU30が実行するプログラムは、PCI I/Fコマンド制御部301と、SCSIコマンド制御部302と、下位ディスクアレ制御部303と、ホストI/Fモード制御部305とから構成されている。PCI I/Fコマンド制御部301は、PCI I/F32を介してホスト計算機1からのコマンドを受信するものであり、第1コマンド受信手段である。PCI I/Fコマンド制御部301は、DMA制御部を含んでおり、このDMA制御部は、ホスト計算機1の主記憶メモリ11とディスクアレコントローラ3のキャッシュメモリ36間のデータ転送制御を実行する。SCSIコマンド制御部302は、SCSI33を介してホスト計算機1からのコマンドを受信するものであり、第2コマンド受信手段である。SCSIコマンド制御部302は、DMA制御部を含んでおり、このDMA制御部は、ホスト計算機1の主記憶メモリ11とディスクアレコントローラ3のキャッシュメモリ36間のデータ転送制御を実行する。なお、図1に示す内蔵型においては、SCSIコマンド制御部302は、使用されていない。下位ディスクアレ制御部303は、上位ディスクアレ制御ドライバ103の生成したディスクコマンドやパリティ生成コマンドを実行するものである。ホストI/Fモード制御部305は、ホストI/Fモード選択手段37により選択されたホストI/Fで、ホスト計算機1とコマンドの授受をできるように制御する。図1に示す構成においては、ホストI/Fモード制御部305からの制御信号によって、PCI I/F32がホストI/Fとして選択されており、ホスト計算機1とコマンドの授受をできるように制御する。

【0050】下位ディスクアレ制御部303は、コマンドI/F制御部3031と、ディスクコマンド制御部3032と、パリティ生成コマンド制御部3033と、初期化部3035と、コマンド管理部3036とから構成されている。コマンドI/F制御部3031は、ホストI/Fモード制御部305により決定されたPCI I/F制御部301もしくはSCSI制御部302のどちらか一方からコマンドを受信するものである。ディスクコマンド制御部3032は、ディスクI/F34a、34bのSCSIコントローラ34a1、34b1を制御し、ディスクコマンドを実行するものである。パリティ生成コマンド制御部3033は、パリティ生成回路35を制御し、パリティ生成コマンドを実行するものである。初期化部3035は、ディスクアレコントローラ3の初期化処理を実行するものである。コマンド管理部3036は、受信したコマンドの管理を行うものである。

#### 【0051】〔動作の説明〕

##### (1) ホストI/Fモードの選択

次に、図3及び図4を用いて、本発明の一実施形態による内蔵型ハイブリッドアレ構成のディスクアレ制御

システムにおけるホストI/Fモード選択手段37の構成及び機能について説明する。図3は、本発明の一実施形態によるディスクアレ制御システムのホストI/Fモード選択手段の構成を示す回路図であり、図4は、本発明の一実施形態によるディスクアレ制御システムのホストI/Fモード選択手段の論理図である。

【0052】ホストI/Fモード選択手段37は、例えば、図3に示すような構成を有している。スイッチ371の開閉により、ホストI/Fモード選択手段37の出力信号であるIF\_Mode信号372は、“1”と“0”の2値を持つことができる。

【0053】図4に示すように、ホストI/Fモード制御部305は、IF\_Mode信号372に応じて、ホストI/Fを選択する。即ち、IF\_Mode信号372が“1”の時、ホストI/Fモード制御部305は、ホストI/Fとして、第1ホストI/FであるPCI I/F32を選択し、IF\_Mode信号が“0”の時、ホストI/Fとして、第2ホストI/FであるSCSI33を選択する。図3に示す状態では、IF\_Mode信号372は“1”であるので、PCI I/F32を選択していることを示している。

【0054】以下の説明においては、ホストI/Fモード制御部305は、IF\_Mode信号372に基づいてPCI I/F32を選択し、図1に示したように、ディスクアレコントローラ3がホスト計算機1のPCIコネクタ14に接続している場合について説明する。

【0055】ディスクアレ装置2の電源投入時に、MPU30は初期化を開始する。初期化開始後、MPU30は、メモリ42からプログラムをリードし、下位ディスクアレ制御部303が動作可能なように初期化を実施する。このとき、ホストI/Fモード選択手段37が出力するIF\_Mode372信号は、MPU30に入力され、下位ディスクアレ制御部303のホストI/Fモード制御部305は、選択されたホストI/Fに対応するコマンド受信手段を選択する。

【0056】図3に示す場合においては、PCI I/F32が選択されているので、MPU30は、PCI I/F制御部301を以降使用する。また、コマンドI/F制御部3031をPCI I/F制御部301と接続するよう初期化する。さらに、SCSIコントローラ331、34a1、34b1やパリティ生成回路35等のハードウェアの初期化も実施する。

【0057】また、以降、選択したホストI/Fとは異なるI/Fからコマンドが送信されてもディスクアレコントローラ3は応答しないか、エラー応答を返すか何れかの方法でホスト計算機に以後のコマンド発行を抑止するよう知らせる。

##### 【0058】(2) コントローラの検出

次に、図5を用いて、本発明の一実施形態による内蔵型ハイブリッドアレ構成のディスクアレ制御システム

(10)

特開平10-105347

におけるディスクアレイドコントローラの検出及び初期化処理について説明する。図5は、本発明の一実施形態によるディスクアレイド制御システムにおけるディスクアレイドコントローラ検出処理を説明するフローチャートである。

【0059】図1に示すホスト計算機1に電源が投入されると、図2に示したOS101は、初期化処理の段階で、各周辺機器用のデバイスドライバを組み込み、デバイスドライバは周辺機器の検出、初期化を行う。ディスクアレイド装置2についても同様である。

【0060】OS101は、上位ディスクアレイド制御ドライバ103を組み込む。上位ディスクアレイド制御ドライバ103のコントローラ検出部1035は、図5に示すフローチャートに従い、ディスクアレイドコントローラの検出、初期化処理を行う。

【0061】図5のステップ1000において、上位ディスクアレイド制御ドライバ103がロードされ、処理が開始される。最初に、ステップ1001において、上位ディスクアレイド制御ドライバ103のコントローラ検出部1035は、PCI I/F32を選択する。次に、ステップ1002において、コントローラ検出部1035は、PCIバス13を走査し、ディスクアレイドコントローラ3があるかどうかを検査する。具体的には、PCIバス13に接続する全てのデバイス（コントローラ）のベンダIDとデバイスIDを示すレジスタをリードし、ディスクアレイドコントローラ3のPCI I/Fコントローラ321の持つベンダID、デバイスIDを検査する。

【0062】ステップ1003において、検出の有無を判断し、検出した場合には、ステップ1008において、コントローラ検出部1035は、ディスクアレイドコントローラ3へのコマンド発行等の通信には、PCI I/F32を使用することとし、上位ディスクアレイド制御ドライバ103とPCI I/Fドライバ104とを接続するようコマンドI/F制御部1034を初期化する。また、PCIバス13上にディスクアレイドコントローラ3を検出できなかったときには、ステップ1004において、コントローラ検出部1035は、SCSIを選択する。次に、ステップ1005において、コントローラ検出部1035は、SCSIを走査し、ディスクアレイドコントローラ3があるかどうかを検査する。具体的には、SCSIに接続する全てのデバイス（コントローラ）にコントローラ種別を問かけるコマンド（Inquiryコマンド）をSCSIドライバ経由105で発行し、その返値から、ディスクアレイドコントローラ3が存在するかどうかを特定する。

【0063】ステップ1006において、検出の有無を判断し、検出した場合には、以降、ディスクアレイドコントローラ3へのコマンド発行等の通信にはSCSIを使用することとし、上位ディスクアレイド制御ドライバ10

3とSCSIドライバ105とを接続するようコマンドI/F制御部1034を初期化する。

【0064】なお、ステップ1004、1005、1006における処理は、PCIバス13にSCSIが接続され、このSCSIとディスクアレイドコントローラ3のSCSI33が接続される場合の処理であり、このような接続関係を「外付け型」と称し、「外付け型」の詳細については、図20を用いて後述する。

【0065】ステップ1010において、上位ディスクアレイド制御ドライバ103のコントローラ検出部1035は、上述したように、何れかのホストI/F上にディスクアレイドコントローラ3を検出したならば、ディスクアレイドコントローラの初期化コマンドを生成し、各ホストI/Fに対応するPCI I/Fドライバ104、若しくはSCSIドライバ105を介して、ディスクアレイドコントローラ3にディスクアレイド構成情報リードコマンドを発行する。ディスクアレイドコントローラ3のPCI I/F制御部301若しくはSCSI制御部302は、ディスクアレイド構成情報リードコマンドを受信し、下位ディスクアレイド制御部303の初期化部3035は、ディスク21の特定の領域に格納しているディスクアレイドの構成情報をリードし、ホスト計算機1にディスクアレイド構成情報を転送する。

【0066】次に、ステップ1011において、コントローラ検出部1035は、このディスクアレイド構成情報に基づき、上位ディスクアレイド制御ドライバ103の初期設定を実施する。

【0067】また、もし、ステップ1006において、コントローラ検出部1035が、何れのホストI/Fにもディスクアレイドコントローラ3を検出できなかったときは、ステップ1007において、上位ディスクアレイド制御ドライバはアンロードされる。

【0068】(3) ディスクアレイドの動作  
次に、図1に示すように、ホストI/FとしてPCI I/F32を選択した時のディスクアレイドの動作について説明する。

【0069】(a) リード動作

次に、図6を用いて、本発明の一実施形態による内蔵型ハイブリッドアレイド構成のディスクアレイド制御システムにおけるディスクアレイドのリード動作について説明する。図6は、本発明の一実施形態によるディスクアレイド制御システムにおけるディスクアレイドのリード処理の説明図である。

【0070】アプリケーション100からディスクアレイド装置2へのリード要求が発行されたものとする。リードデータAは、主記憶メモリ11の領域111にリードされる。リードデータAは、ディスクアレイド装置2のストライプサイズより大きく、図6に示すように、ディスク21aのストライプsaに格納されたデータブロックaと、ディスク21bのストライプsbに格納されたデ

(11)

特開平10-105347

ータブロックbと、ディスク21cのストライプscの一部に格納されたデータブロックcとによって構成されているものとする。

【0071】次に、図7～図15を用いて、ディスクアレイのリード動作について順次説明する。

【0072】(i) コマンドリンクの発行  
アプリケーション100が、リードデータAのリード要求を発行する。ファイルシステム102は、このリード要求を受信し、ディスクアレイ21への論理リードコマンドを上位ディスクアレイ制御ドライバ103に発行する。

【0073】次に、図7を用いて、本発明の一実施形態によるディスクアレイ制御システムにおける上位ディスクアレイ制御ドライバ103のコマンドリンクの発行方法について説明する。図7は、本発明の一実施形態によるディスクアレイ制御システムにおける上位ディスクアレイ制御ドライバのコマンドリンクの発行方法について説明するフローチャートである。

【0074】図7のステップ1100において、上位ディスクアレイ制御ドライバ103の論理コマンドアドレス変換部1031は、論理リードコマンドを受信する。ステップ1101において、論理コマンドアドレス変換部1031は、受信した論理リードコマンドの論理アドレスを、ディスクアレイを構成する少なくとも1台以上の個々のディスクのアドレスに変換する。この時点で、図6に示すように、リードデータAがディスク21a、21b、21cのストライプsa、sb、scに格納されるデータブロックa、b、cにより構成されていることがわかる。

【0075】次に、ステップ1102において、キャッシュ管理部1032は、データブロックa、b、c毎に、ディスクアレイコントローラ3のキャッシュメモリ36のヒットミス判定を実施する。図6に示す例においては、データbはヒット、データa、cはミスヒットであったとする。ステップ1103において、ヒット、ミスヒットの判定を行い、ヒット時には、ステップ1104において、キャッシュ管理部1032は、キャッシュメモリ36から主記憶メモリ11にデータを転送するためのディスクアレイコントローラへのコマンドである「キャッシュリードコマンド」を生成する。以下、ディスクアレイコントローラへのコマンドを「コントローラコマンド」と称する。なお、図7のステップ1104に記載した②「キャッシュライトコマンド」は、ライト動作時のコマンドであり、これについては、ライト動作の説明において後述する。

【0076】ミスヒット時には、ステップ1105において、キャッシュ管理部1032は、キャッシュメモリ36にデータをロードする領域を割り当てる。そして、ステップ1106において、キャッシュ管理部1032は、ディスク21からキャッシュメモリ36にロード

後、主記憶メモリ11にデータを転送するための「キャッシュロード&リードコマンド」を生成する。なお、図7のステップ1106に記載した②「キャッシュロード&ライトコマンド」及び③「キャッシュライトコマンド」は、ライト動作時のコマンドであり、これについては、ライト動作の説明において後述する。

【0077】ステップ1107において、キャッシュ管理部1032は、生成したコントローラコマンドについて、同一論理コマンド毎にコマンドリンクを生成し、主記憶メモリ11の特定の領域に格納する。ステップ1108において、キャッシュ管理部1032は、ステップ1102～1107の処理を、論理コマンドの扱うリードデータを構成する全てのデータブロックについて実施する。

【0078】図6に示す例では、データブロックa、cに対応するキャッシュロード&リードコマンドと、データブロックbに対応するキャッシュリードコマンドの3つのコマンドがリンクされ、主記憶メモリ11に格納される。

【0079】以下、このコントローラコマンドをリンクしたコマンド群を「コマンドリンク」と称する。この例の場合、データブロックcは、ストライプcの一部データのみをリードするので、キャッシュロード&リードコマンドには、ディスク21c上のストライプcの先頭アドレスと、データブロックcの先頭アドレスの両者を格納している。

【0080】次に、ステップ1109において、コマンドI/F制御部103は、PCII/Fドライバ104に指示し、コマンドリンクをディスクアレイコントローラ3に発行する。

【0081】ここで、図8を用いて、ホスト計算機1とディスクアレイコントローラ3の間のコマンドリンクの発行の方法について説明する。図8は、本発明の一実施形態によるディスクアレイ制御システムにおけるホスト計算機とディスクアレイコントローラの間のコマンドリンクの発行方法の説明図である。

【0082】ディスクアレイコントローラ3のPCII/Fコントローラ321は、図8に示すように、DMAコントローラ3211と、MailBoxレジスタ3212とDoorBellレジスタ3213の2つのレジスタを備えている。CPU10のPCII/Fドライバ104は、主記憶メモリ11上のコマンドリンク11のアドレスとその識別番号を含む「コマンドリンク発行パケット」を、MailBoxレジスタ3212にライトする(図8(イ))。すると、PCII/Fコントローラ321は、MPU30に割り込みを発行する(図8(ロ))。

【0083】次に、図9を用いて、PCII/Fコントローラ321から割り込みを受信したMPU30の処理について説明する。図9は、本発明の一実施形態によ

(12)

特開平10-105347

るディスクアレイ制御システムにおけるMPUのコマンドリンクの割り込み処理について説明するフローチャートである。

【0084】図9のステップ1200において、MPU30は割り込みを受信する。

【0085】次に、ステップ1201において、MPU30のPCI I/F制御部301は、MailBoxレジスタ3212をリードし、「コマンドリンク発行パケット」が送信されたことを認識し、DoorBellレジスタ3213の「パケット受領ビット」をセットする。ステップ1202において、PCI I/F制御部301は、コマンドリンクをホスト計算機1の主記憶メモリ11からディスクアレイコントローラ3のメモリ42に転送するための「DMAコマンド」を生成する。さらに、ステップ1203において、PCI I/F制御部301は、PCI I/F制御部301が備えているDMA制御部のDMAコマンドキューにキューイングする。ここで、図7のステップ1110に戻り、ホスト計算機1のPCI I/Fドライバ104は、DoorBellレジスタ3213の「パケット受領ビット」がセットされたことを確認する。

【0086】以上の処理によって、コマンドリンクの発行処理を終了する。以降、コマンドリンクがすべて実行終了し、終了割り込みを受信するまで、CPU10はその他の処理を実行することができる。

【0087】次に、図10～図14を用いて、ディスクアレイコントローラ3の内部処理について説明する。

【0088】(ii) コマンドリンクの受信

次に、図6及び図10を用いて、DMAコマンドキューにコマンドがキューイングされた時の、ディスクアレイコントローラ3のMPU30のPCI I/F制御部301の中のDMA制御部の処理について説明する。図10は、本発明の一実施形態によるディスクアレイ制御システムにおけるMPUのPCI I/F制御部の中のDMA制御部の処理について説明するフローチャートである。

【0089】ステップ1210において、PCI I/F制御部301の中のDMA制御部は、DMAコマンドキューの先頭のDMAコマンドを取り出す。次に、ステップ1211において、PCI I/F制御部301の中のDMA制御部は、PCI I/Fコントローラ321のDMAコントローラ3211にDMAコマンドを発行し、DMAを起動する(図8(ハ))。DMAが起動すると、コマンドリンク111が、ホスト計算機1の主記憶メモリ11からディスクコントローラ3のメモリ42にDMA転送される(図8(ニ))。PCI I/F制御部301の中のDMA制御部は、DMA転送が終了するまでの間スリープし、MPU30は他の処理を実行する事ができる。

【0090】次に、DMA転送が終了すると、PCI

I/Fコントローラ321のDMAコントローラ3211は、割り込みを発行し(図8(ホ))、ステップ1212において、この割り込みをMPU30は受信し、PCI I/F制御部301の中のDMA制御部は動作を再開する。次に、ステップ1213の中のステップ1213aにおいて、PCI I/F制御部301の中のDMA制御部は、下位ディスクアレイ制御部303のコマンドI/F制御部3031のホストコマンドキューに、受信したコマンドリンクをキューイングし、DMAコマンドキューから終了したDMAコマンドを削除する。なお、ステップ1213bは、データ転送時の処理であり、これについては、後述する。

【0091】ステップ1214において、PCI I/F制御部301の中のDMA制御部は、DMAコマンドキューにまだコマンドがあるか否かを判断し、DMAコマンドキューにまだコマンドがあるならば、ステップ1210に戻り、ステップ1210～1213を繰り返すものである。

【0092】(iii) コントローラコマンド(コマンドリンク)の起動処理

次に、図11及び図12を用いて、コマンドI/F制御部3031のホストコマンドキューにコマンドリンクがキューイングされた時の、下位ディスクアレイ制御部303における処理について説明する。図11及び図12は、本発明の一実施形態によるディスクアレイ制御システムにおけるMPUの下位ディスクアレイ制御部の処理について説明するフローチャートである。なお、図11におけるX、Y、Zは、それぞれ、図12のX、Y、Zに連結される。

【0093】ステップ1220において、コマンドI/F制御部3031のホストコマンドキューにコマンドリンクがキューイングされると、下位ディスクアレイ制御部303のコマンド管理部3036は、ホストコマンドキューからコマンドリンクを一つ取り出す。

【0094】次に、ステップ1221において、下位ディスクアレイ制御部303のコマンド管理部3036は、そのコマンドリンクに登録されたコントローラコマンドを一つ取り出し、ステップ1221において、そのコントローラコマンドを解析する。解析されたコントローラコマンドの種類によって、ステップ1227の「キャッシュリードコマンド」と、ステップ1223の「キャッシュロード&リードコマンド」と、ステップ1234の「キャッシュライトコマンド」と、ステップ1230の「キャッシュロード&ライトコマンド」と、ステップ1236の「キャッシュロード&デステージコマンド」と、ステップ1242の「パリティ生成&デステージコマンド」と、ステップ1246の「キャッシュデステージコマンド」との各処理に分かれる。これらの各処理の詳細については、後述する。

【0095】ステップ1249において、コマンド管理

(13)

特開平10-105347

部3036は、全コントローラコマンドの処理が終了したか否かを判断し、全コントローラコマンドの処理が終了するまで、各コントローラコマンド毎の処理が終了すると、次のコントローラコマンドを取り出し、以下コマンドリンクに登録されたすべてのコントローラコマンドの起動処理を行う。

【0096】全コントローラコマンドの処理が終了すると、ステップ1250において、コマンド管理部3036は、ホストコマンドキューから次のコマンドリンクを取り出し、上記の起動処理を繰り返し、すべてのキューイングされたコマンドリンクの起動が終了するとコマンド管理部3036は処理を終了する。

【0097】ここで、図6に示した例では、コントローラコマンドとして、データブロックa、cに対する「キャッシュロード&リードコマンド」と、データブロックbに対する「キャッシュリードコマンド」が使用される。データブロックa、b、cの順にコントローラコマンドの起動処理が行われる。以下、この2つのコマンドについて、コマンド管理部3036の動作を図11を用いて説明する。

【0098】図11のステップ1223において、コマンド管理部3036が、「キャッシュロード&リードコマンド」と判断すると、ステップ1224に進む。ステップ1224において、コマンド管理部3036は、当該ストライプの上位ディスクアレイ制御ドライバ103の指定したキャッシュメモリ36のアドレスに、ディスク21からデータをロードするためのディスクリードコマンドを生成する。なお、本実施形態では、必ずキャッシュへのロードは、ストライプ単位であるとする。よって、データブロックcは、アプリケーションからのリード要求はストライプの一部ではあるが、上位ディスクアレイ制御ドライバの指示に基づき、ディスク21cからキャッシュメモリ36へのロードは、ストライプsc全部をロードすることになる。

【0099】ステップ1225において、コマンド管理部3036は、生成したディスクコマンドを、ディスクコマンド制御部3032のディスクコマンドキューにキューイングする。さらに、ステップ1226において、コマンド管理部3036は、このコントローラコマンドを「コマンドペンディング状態」に設定し、このコントローラコマンドの処理は一旦中断し、コントローラコマンドの終了を待つものである。次に、コマンド管理部3036は、図12のステップ1249に進み、次のコントローラコマンドの処理を続行する。

【0100】ここで、「コマンドペンディング状態」とは、「キャッシュロード&リードコマンド」のように、

(1) 第1フェーズ：ディスクからキャッシュメモリへのデータのロードと、(2) 第2フェーズ：キャッシュメモリから主記憶メモリへのデータのリードの2つの処理を逐次的に行う必要がある場合、第1フェーズを起動

後、一旦、このコントローラコマンドを「コマンドペンディング状態」に設定し、第1フェーズが終了後、第2フェーズの処理を継続するよう制御するためのものである。「コマンドペンディング状態」のコントローラコマンドの第1フェーズが終了した後に、図11及び図12に点線で示すコントローラコマンド種の処理を、第2フェーズとして実行する。同図の他のコントローラコマンド処理の点線はすべて第1フェーズ終了後の第2フェーズ、若しくは、第2フェーズ終了後の第3フェーズを示している。

【0101】「キャッシュロード&リードコマンド」の例においては、ディスクコマンド終了後は、図11の点線に従い、以下に説明する「キャッシュリードコマンド」の処理を実施する。

【0102】図11のステップ1227において、コマンド管理部3036が、「キャッシュリードコマンド」と判断すると、ステップ1228に進む。ステップ1228において、コマンド管理部3036は、キャッシュメモリ36のデータを主記憶メモリ11にDMA転送するためのキャッシュリードDMAコマンドを生成する。そして、ステップ1229において、コマンド管理部3036は、キャッシュリードDMAコマンドをDMAコマンドキューにキューイングし、キャッシュリードコマンドの終了を待つものである。

【0103】なお、ステップ1234の「キャッシュライトコマンド」と、ステップ1230の「キャッシュロード&ライトコマンド」と、ステップ1236の「キャッシュロード&デステージコマンド」と、ステップ1242の「パリティ生成&デステージコマンド」と、ステップ1246の「キャッシュデステージコマンド」との各処理の詳細については、ライト処理において後述する。

【0104】(iv) ディスクコマンドの処理

次に、図13を用いて、ディスクコマンドキューにコマンドがキューイングされた時の、下位ディスクアレイ制御部303のディスクコマンド制御部3032における処理について説明する。図13は、本発明の一実施形態によるディスクアレイ制御システムにおけるMPUの下位ディスクアレイ制御部のディスクコマンド制御部の処理について説明するフローチャートである。

【0105】ステップ1260において、ディスクコマンド制御部3032は、ディスクコマンドキューから一つのディスクコマンドを取り出す。ステップ1261において、ディスクコマンド制御部3032は、取り出したディスクコマンドをSCSIコントローラ34に発行して、ディスクコマンドを起動する。ステップ1262において、ディスクコマンド制御部3032は、まだディスクコマンドキューにディスクコマンドがあるか否かを判断し、ディスクコマンドキューにディスクコマンドが残っており、未動作のディスクがあったり、1台のデ

(14)

特開平10-105347

ディスクが複数のコマンドを受信できるときには、ディスクコマンドの起動を繰り返す。ディスクコマンドが全て無くなる、若しくはそれ以上の起動ができない時には、ディスクコマンド制御部3032は、処理を停止し、スリープする。この間、MPU30は、他の処理を実行できる。

【0106】SCSIコントローラ34は、内部にDMAコントローラを備えており、DMAコントローラが、キャッシュメモリ36とSCSIコントローラ34間のデータ転送を実行する。

【0107】データ転送が終了すると、SCSIコントローラ34は終了割り込みをMPU30に発行する。ステップ1263において、MPU30は、終了割り込みを受信し、ディスクコマンド制御部3032は、処理を再開する。

【0108】ステップ1264において、ディスクコマンド制御部3032は、終了報告をコマンド管理部3036に対し行う。ステップ1265において、ディスクコマンド制御部3032は、終了したディスクコマンドをディスクコマンドキューから削除する。

【0109】ステップ1266において、ディスクコマンド制御部3032は、ディスクコマンドキューに未起動のコマンドがあるか否かを判断し、ディスクコマンドキューに未起動のコマンドがある場合には、上記の起動処理を実行し、もしなければ処理を終了する。

【0110】(v) DMAコマンドの処理  
キャッシュメモリ36と主記憶メモリ11間のデータのDMA転送は、上述した(i)コマンドリンクの受信と同様、PCI I/Fコントローラ321のDMAコントローラ3211で実行する。

【0111】処理の手順は、上述した(i)コマンドリンクの受信とはほぼ同様であるが、相違点は、図10のステップ1213aの代わりに、ステップ1213bにおいて、データ転送終了時には、コマンド管理部3036に終了報告を行うことと、転送の対象がメモリ42ではなくキャッシュメモリ36であることと、転送の方向が双方向であることだけである。

【0112】(vi) コントローラコマンド(コマンドリンク)の終了処理

次に、図14を用いて、上述した(iv)ディスクコマンドの終了報告、(v)DMAコマンドの終了報告、および下記するパリティ生成コマンドの終了報告を受信した時の、下位ディスクアレイ制御部303のディスクコマンド制御部3032における処理について説明する。図14は、本発明の一実施形態によるディスクアレイ制御システムにおけるMPUの下位ディスクアレイ制御部のディスクコマンド制御部の終了報告受信時の処理について説明するフローチャートである。

【0113】ステップ1251において、ディスクコマンド制御部3032は、コマンド終了報告を受信する。

次に、ステップ1252において、ディスクコマンド制御部3032は、そのコマンドに該当するコントローラコマンドが「コマンドペンディング状態」であるか否かを判断し、そのコマンドに該当するコントローラコマンドが「コマンドペンディング状態」であるならば、ステップ1253に進み、そうでないならば、ステップ1255に進む。

【0114】ステップ1253において、ディスクコマンド制御部3032は、もし、そのフェーズが最終フェーズであるならば、起動に先立ち、コマンドのペンディングフラグを変更して、「コマンドペンディング状態」を解除する。即ち、「キャッシュロード&リードコマンド」の処理において、図11のステップ1226によるコマンドペンディング状態であるとする、次の「キャッシュリードコマンド」は、第2フェーズであり、最終フェーズであるため、「コマンドペンディング状態」を解除する。しかしながら、第2フェーズに続いて第3フェーズがある場合には、「コマンドペンディング状態」の解除は行わない。

【0115】次に、ステップ1254において、ディスクコマンド制御部3032は、引き続き、コントローラコマンドの第2フェーズ、若しくは第3フェーズを起動する。即ち、図11の(A)からステップ1222に戻り、ステップ1249の前の(B)までの処理を行う。例えば、「キャッシュロード&リードコマンド」の処理において、図11のステップ1226によるコマンドペンディング状態であるとする、ステップ1227の「キャッシュリードコマンド」を起動する。

【0116】コントローラコマンドが「コマンドペンディング状態」でないときには、ステップ1255において、ディスクコマンド制御部3032は、そのコントローラコマンドの終了でコマンドリンクに登録されたコントローラコマンドがすべて終了したか否かを判断する。

【0117】そのコントローラコマンドの終了でコマンドリンクに登録されたコントローラコマンドがすべて終了したならば、ステップ1256において、ディスクコマンド制御部3032は、コマンドI/F制御部3031に通知し、ホストコマンドキューから当該コマンドリンクを削除する。

【0118】次に、ステップ1257において、PCI I/F制御部301は、PCI I/Fコントローラ321内部のDoorBellレジスタ3213に終了ステータスとコマンドリンクの識別番号をライトする(図8(ヘ))。この動作により、PCI I/Fコントローラ32は、CPU10に割り込みを発行する(図8(ト))。

【0119】次に、図15を用いて、終了報告を受信した時の、ホスト計算機1のPCI I/Fドライバ104における処理について説明する。図15は、本発明の一実施形態によるディスクアレイ制御システムにおけるホ



(15)

特開平10-105347

スト計算機のPCI I/Fドライバにおける終了報告処理について説明するフローチャートである。

【0120】ステップ1111において、ホスト計算機1のPCI I/Fドライバ104は、PCI I/Fコントローラ32からCPU10への割り込みを受信する。ステップ1112において、PCI I/Fドライバ104は、DoorBellレジスタ3213をリードし、ステップ1113において、コマンドリンクの識別番号を獲得し、ステップ1114において、終了ステータスを確認する。次に、ステップ1115において、PCI I/Fドライバ104は、コマンドリンク識別番号と終了ステータスを、上位ディスクアレイ制御ドライバ103の論理コマンドアドレス変換部1031にコマンドI/F1034経由で報告する。論理コマンドアドレス変換部1031は、コマンドリンクの終了処理を行い、ディスクアレイのリードコマンドの終了報告をファイルシステム102に報告する。

【0121】さらに、OS101は、アプリケーション100にディスクアレイのリードコマンドの終了を報告し、コントローラコマンドの全処理が終了する。

【0122】(b) ライト動作

次に、図16を用いて、本発明の一実施形態による内蔵型ハイブリッドアレイ構成のディスクアレイ制御システムにおけるディスクアレイのライト動作について説明する。図16は、本発明の一実施形態によるディスクアレイ制御システムにおけるディスクアレイのライト処理の説明図である。

【0123】ここで、アプリケーション100からディスクアレイ装置2へのライト要求が発行されたとする。主記憶メモリ11の領域111のライトデータBをライトする。ライトデータBは、ディスクアレイ装置2のストライプサイズより大きく、図16のように、ディスクアレイ21a, 21b, 21cの異なる3つのストライプsa, sb, scにまたがるデータブロックa, b, cにより構成されたとする。

【0124】本実施形態では、ディスクアレイへのデータのライトは、ディスクアレイコントローラ3のキャッシュメモリ36にデータをライトすることで処理を終了するものとする。また、キャッシュメモリ36からディスクへの書き戻しは、非同期的に実施し、これを「デステージ動作」と称し、後述する。

【0125】ライト動作の処理フローは、上述した「(a) リード動作」とほとんど同一であるので、相違点についてのみ説明する。

【0126】(i) コマンドリンクの発行  
アプリケーション100がライトデータBのライト要求を発行する。ファイルシステム102は、このライト要求を受信し、ディスクアレイ21への論理ライトコマンドを上位ディスクアレイ制御ドライバ103に発行する。

【0127】次に、図7に示す処理フローに従って、上位ディスクアレイ制御ドライバ103は、コマンドリンクを発行する。

【0128】図7のステップ1100において、上位ディスクアレイ制御ドライバ103の論理コマンドアドレス変換部1031は、論理ライトコマンドを受信する。

【0129】ステップ1101において、論理コマンドアドレス変換部1031は、受信した論理ライトコマンドの論理アドレスを、ディスクアレイを構成する少なくとも1台以上の個々のディスクのアドレスに変換する。上位ディスクアレイ制御ドライバ103の論理コマンドアドレス変換部1031は、図16に示すように、ライトデータBがディスク21a, 21b, 21cのストライプa, b, cに格納するデータブロックa, b, cにより構成されていることを認識する。

【0130】次に、ステップ1102において、キャッシュ管理部1032は、データブロックa, b, c毎に、ディスクアレイコントローラ3のキャッシュメモリ36のヒットミス判定を実施する。図16に示す例においては、データbはヒット、データa, cはミスヒットであったとする。

【0131】ステップ1103において、ヒット、ミスヒットの判定を行い、ヒット時には、ステップ1104において、キャッシュの割当は不要であるため、キャッシュ管理部1032は、「キャッシュライトコマンド」を生成する。

【0132】ミスヒット時には、ステップ1105において、キャッシュ管理部1032は、キャッシュメモリ36にデータをライトする領域を割り当てる。そして、ステップ1106において、キャッシュ管理部1032は、ストライプの一部にデータをライトするデータブロックcのような場合には、一旦ディスク装置21から当該ストライプを全部キャッシュメモリ36にロードした後、一部データをキャッシュメモリ36上で更新する処理を行う。この処理を行うため、キャッシュ管理部1032は、「キャッシュロード&ライトコマンド」を生成する(1106)。データブロックaのようにストライプ全体をライトする場合は、「キャッシュライトコマンド」を生成する。

【0133】なお、キャッシュメモリ36へのデータブロックのライトによりキャッシュメモリ36とディスクのデータの不整合が発生する。これを識別するため、当該ストライプのキャッシュには「Dirty」の識別子を設定する。なお、後のデステージ処理において、キャッシュメモリ36とディスク21のデータが一致した際には、この識別子を「Clean」に設定する。

【0134】ステップ1107において、キャッシュ管理部1032は、生成したコントローラコマンドについて、同一論理コマンド毎にコマンドリンクを生成し、主記憶メモリの特定の領域に格納する。

(16)

特開平10-105347

【0135】ステップ1108において、キャッシュ管理部1032は、ステップ1102～1107の処理を、論理コマンドの扱うライトデータを構成する全てのデータブロックについて実施する。

【0136】図16に示す例では、データブロックaに対応するキャッシュライトコマンドと、データブロックbに対応するキャッシュライトコマンドと、データブロックcに対応するキャッシュロード&ライトコマンドと、の3つのコマンドがリンクされ、キャッシュメモリ36に格納される。

【0137】この例の場合、データブロックcは、ストライプcの一部データのみをライトするので、キャッシュロード&ライトコマンドには、ディスク21c上のストライプcの先頭アドレスと、データブロックcの先頭アドレスの両者を格納している。

【0138】次に、ステップ1109において、コマンドI/F制御部103は、PCII/Fドライバ104に指示し、コマンドリンクをディスクアレイコントローラ3に発行する。

【0139】以下、「コマンドリンク発行パケット」のディスクアレイコントローラへの送信は、上述した「(a) リード動作」の「(i) コマンドリンク発行パケット」と同様である。

【0140】(ii) コマンドリンクの受信

「(a) リード動作」の「(ii) コマンドリンクの受信」と同様である。

【0141】(iii) コントローラコマンド(コマンドリンク)の起動処理

次に、図11及び図12を用いて、コマンドI/F制御部3031のホストコマンドキューにコマンドリンクがキューイングされた時の、下位ディスクアレイ制御部303における処理について説明する。

【0142】「(a) リード動作」と同様に、コマンド管理部3036は、ステップ1220において、ホストコマンドキューからコマンドリンクを一つ取り出し、ステップ1221において、そのコマンドリンクに登録されたコントローラコマンドを一つ取り出し、ステップ1221において、コントローラコマンドを解析する。

【0143】図16の例では、コントローラコマンドとして、データブロックcに対する「キャッシュロード&ライトコマンド」と、データブロックa、bに対する「キャッシュライトコマンド」が使用される。データブロックa、b、cの順にコントローラコマンドの起動処理を行う。以下、この2つのコマンドについてコマンド管理部の動作を図11及び図12を用いて説明する。

【0144】図11のステップ1230において、コマンド管理部3036が、「キャッシュロード&ライトコマンド」と判断すると、ステップ1231に進む。

【0145】ステップ1231において、コマンド管理部3036は、第1フェーズの処理として、当該ストライプ

の全データを上位ディスクアレイ制御ドライバの指定したキャッシュメモリ36のアドレスにディスク21からデータをロードするためのディスクロードコマンドを生成する。なお、本実施形態では、必ずキャッシュへのロードは、ストライプ単位であるとする。よって、データブロックcは、アプリケーションからのリード要求はストライプの一部ではあるが、上位ディスクアレイ制御ドライバの指示に基づき、ディスク21cからキャッシュメモリ36へのロードは、ストライプsc全部をロードすることになる。

【0146】ステップ1232において、コマンド管理部3036は生成したディスクコマンドをディスクコマンド制御部3032のディスクコマンドキューにキューイングする。

【0147】さらに、ステップ1233において、コマンド管理部3036は、このコントローラコマンドを「コマンドペンディング状態」に設定し、このコントローラコマンドの処理は一旦中断し、コントローラコマンドの終了を待つものである。次に、コマンド管理部3036は、図12のステップ1249に進み、次のコントローラコマンドの処理を続行する。ディスクロードコマンド終了後は、同図の点線に従い、以下に示すキャッシュライトコマンドの処理を実施する。

【0148】図11のステップ1234において、コマンド管理部3036が、「キャッシュライトコマンド」と判断すると、ステップ1235に進む。

【0149】ステップ1235において、コマンド管理部3036は、ディスクからデータをキャッシュメモリ36にロードする必要なく、主記憶メモリ11のデータキャッシュメモリ36にDMA転送するためのキャッシュライトDMAコマンドを生成する。

【0150】そして、ステップ1236において、コマンド管理部3036は、キャッシュライトDMAコマンドをDMAコマンドキューにキューイングし、キャッシュライトコマンドの終了を待つものである。

【0151】(iv) ディスクコマンドの処理

「(a) リード動作」の「(iv) ディスクコマンドの処理」と同様である。

【0152】(v) DMAコマンドの処理

「(a) リード動作」の「(v) DMAコマンドの処理」と同様である。

【0153】(vi) コントローラコマンド(コマンドリンク)の終了処理

「(a) リード動作」の「(vi) コントローラコマンド(コマンドリンク)の終了処理」と同様である。「コマンドペンディング状態」のコントローラコマンドは、第2フェーズの処理を実施し、コマンドリンクに登録された全てのコントローラコマンドの処理を実行し終わり、アプリケーションに「(a) リード動作」と同様に終了報告がなされれば、この論理ライトコマンドの処理

(17)

特開平10-105347

はすべて終了する。

【0154】(c) デステージ動作

本実施形態においては、上述したように、データのライト処理は、キャッシュメモリ36に対して実行し、ディスクへの書き戻しは、この処理から遅延させて非同期に実行する。この書き戻し処理を、「デステージ処理」と称する。

【0155】デステージ処理の起動は、ホスト計算機1の上位ディスクアレイ制御ドライバ103が行う。キャッシュ管理部1032は、キャッシュの利用状況を考慮し、例えば、ある一定以上のキャッシュに「Dirty」なブロックが格納されたときや、ライトが行われ、ある一定時間経過したとき等に、キャッシュの「Dirty」ブロックのデステージ処理を起動する。

【0156】デステージ処理は、処理効率をあげるため、まとめ処理を行う等、様々なアルゴリズムを実装できるが、本実施形態では、2つの単純な例で説明する。それ以上の複雑な実装も、その応用として同様に実現できるものである。

【0157】ここで、図17及び図18を用いて、デステージ動作について説明する。

【0158】(i) コマンドリンクの発行

ここで、図17を用いて、1つのDirtyブロックC(新データストライプ)のデステージ処理の動作について説明する。図17は、本発明の一実施形態によるディスクアレイ制御システムにおけるディスクアレイの第1のデステージ処理の説明図である。

【0159】本例においては、1つのDirtyブロックC(新データストライプ)367をデステージする場合について説明する。

【0160】上位ディスクアレイ制御ドライバ103のキャッシュ管理部1032は、DirtyブロックCのデステージを行うことを決定すると、上記同様、コントローラコマンドを生成する。

【0161】この例では、新パリティストライプ367を生成するために必要な旧データストライプ368と旧パリティストライプ370がキャッシュメモリ36上に存在しないとすると、(1)これらのストライプ368、370をリードし、(2)新データストライプであるDirtyブロックC367の3者の排他的論理和(以下、「XOR」と略記する)を演算することで新パリティストライプ369を生成し、(3)DirtyブロックC367および新パリティストライプ369の両者をディスク21にライトする、という3フェーズを実行する必要がある。

【0162】この処理のためのコントローラコマンドとして、「キャッシュロード&デステージコマンド」を生成する。「キャッシュロード&デステージコマンド」については、図12を用いて後述する。「キャッシュロード&デステージコマンド」には、上記(1)から(3)

の処理に必要な、キャッシュメモリのアドレス情報、ディスクのアドレス情報、コントローラコマンドの識別番号等の必要な情報をすべて格納する。

【0163】また、図18を用いて、ディスクアレイの同一パリティグループを構成する全てのデータブロックをまとめてデステージするデステージ処理の動作について説明する。図18は、本発明の一実施形態によるディスクアレイ制御システムにおけるディスクアレイの第2のデステージ処理の説明図である。

【0164】本例においては、ディスクアレイのディスク数は5台で、パリティグループを4つのデータと1つのパリティで構成する4D+1P構成としている。

【0165】上位ディスクアレイ制御ドライバ103のキャッシュ管理部1032は、Dirtyブロック(新データストライプ)D0、D1、D2、D3(371、372、373、374)のデステージを行うことを決定すると、新パリティストライプPは、 $P = D0 + D1 + D2 + D3$  (ただし、+は、XOR演算を示す。)

として生成できるので、(1)パリティストライプPの生成、(2)全新データおよび新パリティストライプのデステージ、の2フェーズを実行する必要がある。

【0166】この処理のためのコントローラコマンドとして、「キャッシュデステージコマンド」を生成する。「キャッシュデステージコマンド」については、図12を用いて後述する。「キャッシュデステージコマンド」には、上記(1)から(2)の処理に必要な、キャッシュメモリのアドレス情報、ディスクのアドレス情報、コントローラコマンドの識別番号等の必要な情報をすべて格納する。

【0167】以上の処理をすべて、もしくは一部の対象Dirtyブロックに対し実施し、コマンドリンクを生成し、コマンド発行パケットを上記同様ディスクアレイコントローラ3に送信する。

【0168】(ii) コマンドリンクの受信

「(a) リード動作」の「(ii) コマンドリンクの受信」と同様である。

【0169】(iii) コントローラコマンド(コマンドリンク)の起動処理

図11に示すように、「(a) リード動作」と同様に、コマンド管理部3036は、ステップ1220において、ホストコマンドキューからコマンドリンクを一つ取り出し、ステップ1221において、そのコマンドリンクに登録されたコントローラコマンドを一つ取り出し、ステップ1222において、コントローラコマンドを解析する。

【0170】図17の例では、コントローラコマンドとして、上述したように、「キャッシュロード&デステージコマンド」が使用される。また、図18の例では、コントローラコマンドとして、上述したように、「パリティ

(18)

特開平10-105347

ィ生成&デステージコマンド」が使用される。

【0171】次に、図12を用いて、これらのコントローラコマンドの処理におけるコマンド管理部3036の動作について説明する。

【0172】図12のステップ1236において、コマンド管理部3036が、「キャッシュロード&デステージコマンド」と判断すると、ステップ1237に進む。ステップ1237において、コマンド管理部3036は、第1フェーズの処理として、旧データストライプ368と、旧パリティストライプ370の両者をディスク21cからキャッシュメモリ36にロードするためのディスクリードコマンドを生成する。ステップ1238において、コマンド管理部3036は、生成したディスクコマンドを、ディスクコマンド制御部3032のディスクコマンドキューにキューイングする。

【0173】さらに、ステップ1239において、コマンド管理部3036は、このコントローラコマンドを「コマンドペンディング状態」に設定し、このコントローラコマンドの処理は一旦中断し、コントローラコマンドの終了を待つものである。ディスクコマンド終了後は、第2フェーズとして、同図の点線に従い、以下に示す「パリティ生成&デステージコマンド」の処理を実施する。

【0174】図12のステップ1242において、コマンド管理部3036が、「パリティ生成&デステージコマンド」と判断すると、ステップ1243に進む。

【0175】ステップ1243において、コマンド管理部3036は、第1フェーズの処理として（キャッシュロード&デステージコマンドから引き継いだときは、第2フェーズの処理として）、新パリティストライプの生成のために必要なデータストライプのXOR演算を実行するための「パリティ生成コマンド」を生成する。図17に示した第1の例では、新データストライプ367と、旧データストライプ368と、旧パリティ370とをXOR演算して、新パリティ369を生成するための「パリティ生成コマンド」を生成する。また、図18に示した例では、新データストライプD0、D1、D2、D3（371、372、373、374）のXOR演算を実行して、新パリティ375を生成するための「パリティ生成コマンド」を生成する。

【0176】ステップ1244において、コマンド管理部3036は、生成したパリティ生成コマンドをパリティ生成コマンド制御部3033のパリティ生成コマンドキュー30331にキューイングする。

【0177】さらに、ステップ1245において、コマンド管理部3036は、このコントローラコマンドを「コマンドペンディング状態」に設定し、このコントローラコマンドの処理は一旦中断し、コントローラコマンドの終了を待つものである。ディスクコマンド終了後は、第2フェーズとして、同図の点線に従い、以下に示

す「キャッシュデステージコマンド」の処理を実施する。

【0178】「キャッシュデステージコマンド」は、ホスト計算機1の上位ディスクアレイ制御ドライバ103からは発行されないコントローラコマンドである。キャッシュデステージコマンドは、キャッシュロード&デステージコマンド、及びパリティ生成&デステージコマンドから引き継いで、第2、若しくは第3フェーズの処理として実行され、コントローラコマンドに登録されたDirtyな新データストライプもしくは生成した新パリティストライプのディスクへのライトを行うものである。

【0179】図12のステップ1246において、コマンド管理部3036が、「キャッシュデステージコマンド」と判断すると、ステップ1247に進む。

【0180】ステップ1247において、コマンド管理部3036は、ディスクライトコマンドを生成する。図17の例では、新データストライプC367および新パリティストライプ369をデステージするため、これら全てのディスクライトコマンドを生成する。また、図18に示す例では、新データストライプD0、D1、D2、D3（371、372、373、374）および新パリティストライプP375をデステージするため、これら全てのディスクライトコマンドを生成する。

【0181】ステップ1248において、コマンド管理部3036は、これらのディスクライトコマンドをディスクコマンドキューにキューイングし、全ディスクライトコマンドの終了を待つものである。

【0182】(iv) ディスクコマンドの処理  
「(a) リード動作」の「(iv) ディスクコマンド」と同様である。

【0183】(v) パリティ生成コマンドの処理  
次に、図19を用いて、パリティ生成コマンドキュー30331にコマンドがキューイングされた時の、下位ディスクアレイ制御部303のパリティ生成コマンド制御部3033における処理について説明する。図19は、本発明の一実施形態によるディスクアレイ制御システムにおけるMPUの下位ディスクアレイ制御部のパリティ生成コマンド制御部の処理について説明するフローチャートである。

【0184】ステップ1270において、パリティ生成コマンド制御部3033は、パリティ生成コマンドキュー30331から一つのコマンドを取り出す。ステップ1271において、パリティ生成コマンド制御部3033は、パリティ生成コマンドを起動する。パリティ生成コマンド制御部3033は、処理を停止し、スリープする。この間、MPU30は他の処理を実行できる。パリティ生成回路35は、キャッシュメモリ36からXOR演算対象データをリードし、演算結果を新パリティとしてキャッシュメモリ36にライトする。

(19)

特開平10-105347

【0185】パリティ演算が終了すると、パリティ生成回路35は終了割り込みをMPU30に発行する。ステップ1272において、MPU30は、終了割り込みを受信する。そして、パリティ生成コマンド制御部3033が処理を再開する。ステップ1273において、パリティ生成コマンド制御部3033は、終了報告をコマンド管理部3036に対し行う。ステップ1274において、パリティ生成コマンド制御部3033は、この終了したパリティ生成コマンドを、パリティ生成コマンドキュー30331から削除する。

【0186】ステップ1274において、パリティ生成コマンド制御部3033は、パリティ生成コマンドキュー30331に未起動のコマンドがあるか否かを判断し、パリティ生成コマンドキュー30331に未起動のコマンドがある場合には、上記の起動処理を実行し、なければ処理を終了する。

【0187】(vi) コントローラコマンド(コマンドリンク)の終了処理

「(a) リード動作」の「(vi) コントローラコマンド(コマンドリンク)の終了処理」と同様である。

【0188】「コマンドペンディング状態」のコントローラコマンドは、第2、第3フェーズの処理を実施し、コマンドリンクに登録された全てのコントローラコマンドの処理を実行し終わったなら、上記「(a) リード動作」と同様の方法で、上位ディスクアレイ制御ドライバ103のキャッシュ管理部1032に終了報告を行い、キャッシュ管理部1032は、キャッシュメモリ36とディスク21のデータが一致したので「Dirty」識別子を「Clean」に設定する。以上でデステージ処理は終了する。

【0189】(d) 縮退動作

ディスクアレイを構成するディスク21の内の1つのディスクが故障して、データストライプが失われた場合、パリティストライプから故障ディスクのデータを再現することができる。この動作を、「縮退動作」と称する。

【0190】図18に示したような4D+1P構成のディスクアレイの場合、ディスク21a, 21b, 21c, 21d, 21eのうち、例えば、ディスク21aが故障したとすると、各々に格納されるデータストライプD0, D1, D2, D3のうち、データストライプD0が失われる。しかしながら、データストライプD0は、パリティストライプPを用い、以下の式に基づき再現できる。

【0191】 $D0 = D1 + D2 + D3 + P$  (但し、+はXOR演算を表す。)

また、ディスクアレイを構成するディスク21の内の1つのディスクが故障した状態で、新規なデータストライプD0newをライトするときは、新パリティP1を生成する必要がある。そこで、以下の式に基づき、ディスク21b, 21c, 21dからデータストライプD1,

D2, D3をリードし、新規なデータストライプD0newとのXOR演算の実行により、新パリティP1を生成し、新パリティP1のみをディスク21eにライトする。

【0192】 $P1 = D0new + D1 + D2 + D3$   
(但し、+はXOR演算を表す。)

これらの縮退動作については、図11及び図12に示す処理フロー中には記載していないが、すべて上述した動作の組み合わせで実現できる。または、新規にコントローラコマンドを用意することもできるが、やはり上記動作の応用として容易に実現できるものである。

【0193】以上の説明したように、ディスクアレイコントローラ3を宿主計算機1に内蔵し、宿主I/FとしてPCI I/F32を用いて、上位ディスクアレイ制御ドライバ103と下位ディスクアレイ制御部303との間のコマンド転送や、主記憶メモリ11とキャッシュメモリ36間のデータ転送を行うことができる。

【0194】このとき、第2宿主I/FであるSCSIは用いていないので、SCSIコントローラ331は、ディスク用のSCSIコントローラとして使用することができる。

【0195】以上説明したように、ディスクアレイ処理を、宿主計算機1のCPU10とディスクコントローラ3のMPU30とで分散した「ハイブリッドアレイ」構成とすることにより、処理負荷の重いディスクアレイの処理に、処理能力の高いCPU10で処理負荷の重い上位ディスクアレイ制御を実行し、処理能力の低いMPU30で処理負荷の軽い下位ディスクアレイ制御を実施することができ、従来のディスクアレイコントローラのMPUが性能ボトルネックになることによる限界性能を越える高性能化を実現することができるものである。

【0196】本実施形態によれば、ディスクアレイコントローラに低価格なMPUを使用したディスクアレイ制御システムにおいて、MPUの能力で性能が制限されることなく高性能なディスクアレイ制御が実現できる。

【0197】次に、図20～図22を用いて、本発明の第2の実施形態によるディスクアレイ制御システムについて説明する。

【0198】[構成の説明] 最初に、図20を用いて、本発明の第2の実施形態による外付け型ハイブリッドアレイ構成のディスクアレイ制御システムの全体構成について説明する。図20は、本発明の第2の実施形態によるディスクアレイ制御システムのハードウェアの全体構成のブロック図である。なお、図1と同一符号は、同一部分を示しており、以下の説明においては、図1との相違点を中心に説明する。

【0199】宿主計算機1は、図1に示した構成に加え、SCSI15を制御するSCSIコントローラ151と、SCSI15を接続するSCSIコネクタ152を備えている。

(20)

特開平10-105347

【0200】ディスクアレイ装置2は、ホスト計算機1とは独立した別の筐体を備える。この構成を、「外付け型」と称する。ディスクアレイ装置2は、図1の構成に加え、電源22と、クロック発生器23と、ホストI/FとしてSCSIを選択するSCSI選択手段25とを備えている。また、ディスクアレイ装置2は、筐体内の温度を監視する温度監視手段281と、ディスクの挿抜を検出する挿抜検出手段282と、LEDの点灯制御を行うLED制御手段283と、電源22やFAN200の故障を検出する故障検出手段284と、温度監視手段281と挿抜検出手段282とLED制御手段283と故障検出手段284を制御し、筐体に発生する全ての異常検出を行う筐体異常検出手段28とを備えている。なお、ディスクアレイ装置2は、さらに、ハードアレイ選択手段29を備えているが、これについては、第3の実施形態において説明する。

【0201】ディスクアレイコントローラ3は、図1の構成に加え、使用するホストI/Fを選択するホストI/Fモード選択手段37aと、筐体の異常を監視する筐体異常監視手段43とを備えている。なお、ディスクアレイコントローラ3は、ディスクアレイモード選択手段44を備えているが、これについては、第3の実施形態において説明する。

【0202】次に、図21を用いて、本発明の第2の実施形態による外付け型ハイブリッドアレイ構成のディスクアレイ制御システムの立体的構成について説明する。図21は、本発明の第2の実施形態によるディスクアレイ制御システムのハードウェアのブロック斜視図である。なお、図20と同一符号は、同一部分を示している。

【0203】ディスクアレイ装置2は、さらに、ディスクアレイコントローラ3を搭載するためのコネクタ24と、バックボード290と、バックボード290とディスクアレイコントローラ3を接続するバックボードI/Fコネクタ251と、筐体内部の空調を行うFAN200と、ディスクや電源やFANの状態を表示するLED213a, b, c..., 223, 201を備えている。

【0204】ディスクアレイ装置2のバックボード290は、カードエッジ31を接続するコネクタ24を備えており、ディスクアレイコントローラ2を接続する。ブラケット27aは、ディスクアレイ装置2の筐体に固定されている。ブラケット27aは、ホスト計算機1とSCSIケーブル17に接続されている。筐体には、ブラケット27aの逆側の横辺を固定するための、支え金具27bが設けられている。

【0205】ディスクアレイコントローラ3は、さらに、ディスクアレイ装置2のバックボード290と接続するバックボードI/Fコネクタ373とPCI I/F32に接続するためのカードエッジ31を備えている。

#### 【0206】[動作の説明]

##### (1) ホストI/Fの選択

図22を用いて、ホストI/Fモード選択手段37aによるホストI/Fの選択について説明する。図22は、本発明の第2の実施形態によるディスクアレイ制御システムのホストI/Fモード選択手段の構成を示す回路図である。

【0207】ディスクアレイコントローラ3のホストI/Fモード選択手段37aは、どちらのホストI/Fを選択しているかを示すIF\_Mode信号372をMPU30に送出する。IF\_Mode信号372は、抵抗で電源にプルアップされている。ホストI/Fモード選択手段37aは、バックボードI/Fコネクタ373を備えている。

【0208】ディスクアレイ装置内部のバックボード290は、SCSI選択手段25を備えている。SCSI選択手段25のバックボードI/Fコネクタ251は、バックボードI/Fコネクタ373と対になるものであり、グランド線に接地してされている。

【0209】図21に示したように、ディスクアレイコントローラ3をディスクアレイ装置2のバックボード290に実装すると、ホストI/Fモード選択手段37aのバックボードI/Fコネクタ373と、SCSI選択手段25のバックボードI/Fコネクタ251は接続する。この結果、IF\_Mode信号372は、信号レベルが“L”になる。一方、もし、両者が接続していないときは、バックボードI/Fコネクタ373がオープンになり、プルアップ抵抗により、IF\_Mode信号372は、信号レベルが“H”になる。従って、バックボード290との接続しているかどうかにより、図4に示したように、IF\_Mode信号372が変化することになる。

【0210】以上のように、ディスクアレイコントローラ3を、ホスト計算機1から独立したディスクアレイ装置2に内蔵している「外付け型」のときは、ホストI/Fとして、SCSI33を選択し、そうでないとき、すなわち、ホスト計算機1に内蔵している「内蔵型」のときは、ホストI/Fとして、PCI I/F32を選択することができる。

【0211】本実施形態においては、ディスクアレイコントローラ3がディスクアレイ装置2に内蔵され、バックボード290に接続しているものとする。

【0212】ディスクアレイ装置2の電源が投入されると、IF\_Mode信号372が信号レベル“L”でMPU30に入力し、MPU30で動作する図2に示したホストI/Fモード制御部305はこれを認識し、図5のステップ1009において、ホストI/Fとして、SCSI33を選択し、初期化部3035は、ディスクアレイコントローラ3の初期化を実施する。

【0213】(2) ディスクアレイコントローラの実

(21)

特開平10-105347

装

次に、図21を用いて、ディスクコントローラ3のディスクアレ装置2への実装の方法について説明する。

【0214】ディスクアレコントローラ3は、PCI I/F32に接続するためのカードエッジ31を備えている。ディスクアレ装置2のバックボード290は、カードエッジ31を接続するコネクタ24を備えており、ディスクアレコントローラ2を接続する。ホストI/Fとして使用するSCSI33は、ディスクアレコントローラ2の横辺に備えるブラケット27aに固定したSCSIコネクタ332を用い、ホスト計算機1とSCSIケーブル17によって接続する。このブラケット27aは、ディスクアレ装置2の筐体に固定されている。また、ディスクアレ装置2の筐体には、ブラケット27aの逆側の横辺を固定するための支え金具27bを設けてある。カードエッジ31用コネクタ24、ブラケット27a、支え金具27bの3点によって、ディスクアレコントローラ3は、ディスクアレ装置2の筐体に固定されている。

【0215】また、ディスクアレコントローラ3への電力や動作クロックの供給は、バックボード290からPCI I/F32のカードエッジ31を介して行われる。

【0216】バックボード290上には、クロック発生器23を備えている。また、ディスクアレ装置2には、電源22を備えている。クロック発生器23が送出するクロック信号231と、電源22が送出する電力を伝達する電源線222は、バックボード290に備えてあり、コネクタ24からPCI I/F32のカードエッジ仕様にあわせ接続する。

【0217】以上のような構成により、ディスクアレコントローラ3をホスト計算機1に内蔵してPCI I/F32をホストI/Fとして使用するときと全く同じ方法で、ディスクアレコントローラ3に電力とクロック信号を供給できる。

【0218】(3) 筐体異常の検出と制御  
ディスクアレ装置2は、ホスト計算機1とは独立しているので、上述したように、ディスクアレ装置2は、専用の電源22を備えている。また、装置内部の空調用のファン200を備えている。また、図21には記載していないが、電源もファンも複数台備えることで耐故障性を持つことができる。

【0219】これらの各部位の故障の有無を、ディスクアレコントローラは監視する必要がある。そこで、ディスクアレ装置2は、故障検出手段284を備えており、故障検出手段284が、電源22、ファン200の異常を検出する。

【0220】また、電源22、ファン200は、冗長構成の場合、オンライン中に交換することができる。そこで、どの電源22、ファン200が故障しているかをデ

ィスクアレ装置2のユーザに視覚的に報告する必要があるので、電源22及びファン200は、それぞれ、LED223、201を備えている。LED制御手段283は、LED223、201の点灯制御を行うものである。

【0221】また、ディスクアレ装置2は、ディスクの故障に対しても耐故障性を備えている。ディスクが故障した際には、縮耐運転を行うことで、ディスクアレの処理を継続可能である。ただし、同一パリティグループを構築するディスクが2台以上同時に故障すると、データを喪失してしまうので、1台のディスクが故障したら速やかにディスクを交換する必要がある。そこで、ディスクアレ装置2のユーザには、ディスク装置21a、21b、21c、21dの状態と故障ディスクを視覚的に報告する必要があるので、ディスク装置21a、21b、21c、21dは、LED213a、…、213eを備えている。上記LED制御手段283は、LED213a、…、213eの点灯制御を行う。

【0222】また、故障したディスクが抜き取られたり、新しいディスクが挿入されたりした際には、この状態変化をディスクアレコントローラ3は検出する必要がある。挿抜検出手段282は、ディスクの挿抜を検出する。

【0223】また、ディスクアレ装置2の内部は何らかの異常により、温度が上昇し、装置全体に悪影響を及ぼす恐れがある。そこで、温度監視手段281は、ディスクアレ装置内部の温度を監視する。

【0224】故障検出手段284、LED制御手段283、挿抜検出手段282、温度監視手段281は、ディスクアレ装置2に備えた筐体異常検出手段28に接続される。筐体異常検出手段28は、ディスクアレコントローラ2が備える筐体監視手段43に接続される。筐体異常検出手段28は、ディスクアレ装置2の内部で発生した上記異常を検出し、筐体監視手段43に報告する。また、筐体監視手段43は、必要なLEDの点灯指示を筐体異常検出手段28経由でLED制御手段283に発行する。

【0225】ディスクアレコントローラ3上の筐体監視手段43と、ディスクアレ装置2の筐体異常検出手段28の接続には、上記バックボードI/Fコネクタ373、251を使用する。これにより、特別なケーブル無しにディスクアレコントローラ3が、筐体の異常情報を検出することができる。

【0226】なお、この筐体異常の検出の方法は、第1実施形態記載のホストI/FとしてPCI I/F32を選択した際にも使用することができるものである。

【0227】(4) ディスクアレ動作  
次に、ディスクアレの動作について説明する。本実施形態においては、ホストI/FとしてPCI I/F32ではなくSCSI33を用いるようにしているので、

(22)

特開平10-105347

ホスト計算機1の上位ディスクアレイ制御ドライバ103で生成したコントローラコマンドをディスクアレイコントローラ3の下位ディスクアレイ制御部303に送信する方法が異なるものである。そこで、コントローラコマンドの送信方法について、以下に説明する。

【0228】コントローラコマンドの種類と生成の方法は、上述した第1の実施形態と同様である。相違する点は、SCSI33を用いる場合には、SCSIのコマンドブロックであるCDB (Command Descriptor Block) として実現することである。また、コマンドリンクは、SCSIの仕様であるリンクコマンド機能を用いて実現できる。この機能は、CDBのLinkビットを"1"にセットすることで、連続してCDBを送信し、これらのリンクされたコマンドが終了した時点で終了報告をまとめて受信する。この機能を用いることで、第1の実施形態と等価なコマンドリンクを実現できる。

【0229】次に、コマンドリンクの発行の手順を説明する。図2に示した上位ディスクアレイ制御ドライバ103のキャッシュ管理部1032は、コントローラコマンドを生成し、主記憶メモリ11にコマンドリンクを生成する。コマンドI/F制御部1034は、SCSIドライバ105にコマンド発行を指示する。SCSIドライバ105は、SCSIコントローラ15のコマンド起動ビットをセットする。SCSIコントローラ15は、インテリジェント型である場合、コマンド起動ビットのセットにより、コマンドリンクをSCSIコントローラ15の内部のDMAコントローラでとりこみ、SCSIバスを駆動して、ディスクアレイコントローラ3に送信する。

【0230】ディスクアレイコントローラ3のSCSIコントローラ331は、このコマンドリンクを受信し、その内部のDMAコントローラでメモリ42に転送し、MPU30に割り込みを発行する。MPU30は、この割り込みを受信し、SCSI制御部302を起動する。SCSI制御部302は、受信したコマンドリンクを取り込み、下位ディスクアレイ制御部303のホストコマンドキューにキューイングする。以下の処理は、上述した第1の実施形態と同様である。

【0231】また、ホスト計算機1の主記憶メモリ11とディスクアレイコントローラ3のキャッシュメモリ36との間のデータ転送は、上述した第1の実施形態では、図8に示したPCI I/Fコントローラ321の有するDMAコントローラ3211が直接両者間を転送したが、本実施形態ではSCSIを用いるため、このような転送ができない。

【0232】そこで、主記憶メモリ11とのキャッシュメモリ36との間のデータ転送の方法について、ディスクアレイのキャッシュリードを例にとり、以下に、説明する。ここでは、キャッシュメモリ36から主記憶メ

モリ11へデータが転送される。キャッシュライト時は、方向が逆になるだけで同様である。

【0233】図2において、ディスク21からデータがキャッシュメモリ36に転送されると、コマンド管理部3036は、SCSI制御部302に通知する。SCSIバスは、コマンドを受信した後、キャッシュメモリ36にデータが準備されるまでの間ディスク接続しており、この間、ホストI/FであるSCSI33は他のコマンドの受信や、他のコマンドのデータ転送に用いることができる。SCSI制御部302は、この通知を受けSCSIバスをリコネクトし、SCSIコントローラ331の内部のDMAコントローラを起動し、キャッシュメモリ36からデータをSCSIコントローラ331内部に転送し、このデータをSCSI33経由でホスト計算機1のSCSIコントローラ15に転送する。ホスト計算機1のSCSIコントローラ15の内部のDMAコントローラは、このデータを受信し、主記憶メモリ11にDMA転送する。

【0234】また、コントローラコマンドの終了報告も、SCSI33のステータスを用いて実現することができる。

【0235】なお、デステージ処理は、ホスト計算機1からコントローラコマンドが送信されるだけで、ディスクアレイ装置2内部で処理が実行されるので、SCSIバス上でデータ転送は行われないので、上述した第1の実施形態と同様である。

【0236】以上の説明したように、ディスクアレイコントローラ3をホスト計算機1に内蔵し、ホストI/FとしてPCI I/F32を用いて、上位ディスクアレイ制御ドライバ103と下位ディスクアレイ制御部303との間のコマンド転送や、主記憶メモリ11とキャッシュメモリ36間のデータ転送を行うことができる。

【0237】このとき、第2ホストI/FであるSCSIは用いていないので、SCSIコントローラ331は、ディスク用のSCSIコントローラとして使用することができる。

【0238】以上説明したように、ディスクアレイ処理を、ホスト計算機1のCPU10とディスクコントローラ3のMPU30とで分散した「ハイブリッドアレイ」構成とすることにより、処理負荷の重いディスクアレイの処理に、処理能力の高いCPU10で処理負荷の重い上位ディスクアレイ制御を実行し、処理能力の低いMPU30で処理負荷の軽い下位ディスクアレイ制御を実施することができ、従来のディスクアレイコントローラのMPUが性能ボトルネックになることによる限界性能を越える高性能化を実現することができるものである。

【0239】また、第1の実施形態及び第2の実施形態において説明したように、同一のディスクアレイコントローラを用いながら、ホストI/FをPCIバス等のホストの内部バスと、SCSIバス等のホスト計算機とデ



(23)

特開平10-105347

ィスクアレイ装置をケーブルで接続する外部バスの両者を選択的に使用することができる。

【0240】また、ディスクアレイ装置2のバックボード290とディスクアレイコントローラ3を接続するコネクタを設けることで、両者を接続した際には自動的にホストI/FとしてSCSIを選択するようにできる。

【0241】本実施形態によれば、ディスクアレイコントローラに低価格なMPUを使用したディスクアレイ制御システムにおいて、MPUの能力で性能が制限されることなく高性能なディスクアレイ制御が実現できる。

【0242】また、内蔵型ハイブリッドアレイ構成のディスクアレイ制御システムと、外付け型ハイブリッドアレイ構成のディスクアレイ制御システムとを同一のディスクアレイコントローラを用いて選択的に構成することができる。

【0243】次に、図23～図24を用いて、本発明の第3の実施形態によるディスクアレイ制御システムについて説明する。上述した第2の実施形態においては、ハイブリッドアレイ構成のディスクアレイ制御システムについて説明したが、本実施形態においては、ディスクアレイコントローラ3を、「ハードアレイ」としても、「ハイブリッドアレイ」としても動作可能とし、この両者を選択的に使用できるよう構成したものである。

【0244】ここで、「ハードアレイ」とは、ディスクアレイコントローラ3においてすべてのディスクアレイ制御を行うもので、本明細書においては特にディスクアレイへの論理アドレスを各ディスクのディスクアドレスへアドレス変換する手段が、ディスクアレイコントローラ3上にあるディスクアレイの制御方法と定義する。

【0245】また、「ハイブリッドアレイ」とは、ディスクアレイコントローラ3とホスト計算機1の両方で分担してディスクアレイ制御を行うもので、本明細書においては特に上記アドレス変換する手段がホスト計算機1上にあり、さらに、ディスクアレイ制御のアドレス変換以外のある部分がディスクアレイコントローラ3上にある、ディスクアレイの制御方法と定義する。

【0246】すなわち、「ハイブリッドアレイ」は、上述した第1、2の実施形態のように、上位ディスクアレイ制御手段(上位ディスクアレイ制御ドライバ)がホスト計算機1上にあるもので、「ハードアレイ」は、上位ディスクアレイ制御部がディスクアレイコントローラ3上にあるものである。

【0247】最初に、図23を用いて、本発明の第3の実施形態によるディスクアレイ制御システムのソフトウェアの全体構成について説明する。図23は、本発明の第3の実施形態によるディスクアレイ制御システムのソフトウェアの全体構成のブロック図である。なお、図2と同一符号は、同一部分を示しており、以下の説明においては、図2との相違点を中心に説明する。

【0248】図23において、ホスト計算機1は、図2

に示すプログラムに加えて、さらに、CPU10で動作するディスクドライバ106を備えている。また、ディスクアレイコントローラ3のMPU30は、図2に示すプログラムに加えて、さらに、第2上位ディスクアレイ制御部304と、ディスクアレイモード制御部306を備えている。

【0249】本実施形態におけるハード構成としては、図1若しくは図20に示したハード構成とすることができる。本実施形態においては、図20に示したハード構成に加えて、さらに、ディスクアレイコントローラは、ディスクアレイモード選択手段44を備え、ディスクアレイ装置は、ハードアレイ選択手段29を備えている。ハードアレイ選択手段29は、図20のSCSI選択手段25と全く同一に構成することができる。

【0250】ディスクアレイモード選択手段44とハードアレイ選択手段29の接続は、図21に示したように、バックボードI/Fコネクタ353、251により行うことができる。

【0251】ディスクドライバ106は、ホスト計算機1に単体のディスクを接続するときに、ファイルシステムからのディスクアクセス要求をディスクへのコマンドに変換するものである。一般に、OS101は、標準でディスクドライバ106を備えている。

【0252】第2上位ディスクアレイ制御部304は、ファイルシステム102の代わりにPCI I/F制御部301やSCSI制御部302からディスクアレイへの論理コマンドを受信するものであり、この機能においては、上位ディスクアレイ制御ドライバ103の機能と同一である。

【0253】ディスクアレイモード選択手段44の構成は、図3もしくは図22に示したホストI/Fモード選択手段と全く同様に構成できる。本実施形態においては、図22に示したホストI/Fモード選択手段と同様に構成したものとする。

【0254】また、ディスクアレイモード選択手段44は、ディスクアレイモードとして、「ハイブリッドアレイモード」と「ハードアレイモード」のいずれを選択するかのDiskArray\_Mode信号をMPU30に送出する。

【0255】ここで、図24を用いて、ディスクアレイモード選択手段44が送出するDiskArray\_Mode信号について説明する。図24は、本発明の第3の実施形態によるディスクアレイ制御システムにおけるディスクアレイコントローラのディスクアレイモード選択手段が送出するDiskArray\_Mode信号の論理図である。

【0256】図24に示すように、DiskArray\_Mode信号が“1”の時には、ディスクアレイモードとして、「ハイブリッドアレイモード」を選択し、DiskArray\_Mode信号が“0”の時には、デ

(24)

特開平10-105347

ィスクアレイモードとして、「ハードアレイモード」のいずれを選択するようにしている。なお、DiskArray\_Mode信号[1:0]は、本発明の第5の実施形態に関するものであり、この点については、後述する。

【0257】ディスクアレイモード制御部306は、ディスクアレイモード選択手段44が送出するDiskArray\_Mode信号に従い、「ハードアレイモード」、「ハイブリッドアレイモード」のどちらでディスクアレイコントローラ3を動作させるかを決定する。

【0258】「ハイブリッドアレイモード」の時には、ディスクアレイモード制御部306は、第2上位ディスクアレイ制御部304をディスエーブルにする。また、ディスクアレイモード制御部306は、下位ディスクアレイ制御部303がコントローラコマンドをPCIもしくはSCSIのホストI/F制御部(PCI/F制御部301、SCSI制御部302)から受信するように制御する。

【0259】「ハードアレイモード」の時には、ディスクアレイモード制御部306は、第2上位ディスクアレイ制御部304をイネーブルにする。また、ディスクアレイモード制御部306は、下位ディスクアレイ制御部303がコントローラコマンドを第2上位ディスクアレイ制御部304から受信するように制御する。

【0260】ホストI/F32,33は、「ハイブリッドモード」、「ハードアレイモード」のいずれが選択されていても、PCI/F32とSCSI33のどちらでも使用できる。即ち、PCI/F32を使用することにより、図1に示したような「内蔵型」の構成とすることができ、また、SCSI33を使用することにより、図20に示したような「内蔵型」の構成とすることができる。

【0261】このようにホストI/Fモードとディスクアレイモードは独立なので、上記のようにディスクアレイモード選択手段44を構成した場合、バックボードI/Fコネクタ353,251は、それぞれに専用な信号線を有する必要がある。

【0262】また、図3に示したホストI/Fモード選択手段37のように、スイッチを用いてディスクアレイモード選択手段44を構成した場合、スイッチはそれぞれ独立に備える必要がある。

【0263】本実施形態においては、図20に示したように、ホストI/FとしてSCSI33が選択されていて、ディスクアレイコントローラ3がディスクアレイコントローラ3のバックボード290に接続しているものとする。ディスクアレイ装置2は、ハードアレイ選択手段29を備えるので、図22に示したSCSI選択手段によるホストI/Fの選択と同様に、ハードアレイ選択手段29により「ハードアレイモード」が選択される。もちろん、ハードアレイ選択手段を切り換え可能に構成

し、「ハードアレイモード」と「ハイブリッドアレイモード」を切り換えることもできる。

【0264】上述したように、「ハードアレイ」と「ハイブリッドアレイ」の何れのディスクアレイでも、同一ディスクアレイコントローラを用い、かつ、ホストI/FもPCIやSCSIによらず構築することができ、高性能が必要な場合には高速なホスト計算機のCPUで上位ディスクアレイ制御を行える「ハイブリッドアレイモード」を選択でき、また、ホスト計算機やOSに非依存なディスクアレイを構築したい際や、CPUの負担を下げたい(CPU負荷率を低減したい)際には、標準のディスクドライバを使用できる「ハードアレイ」を選択することができるものである。このように、構築自由度の高いディスクアレイを構成することができるものである。

【0265】以上説明したように、本実施形態によれば、ホスト計算機のホストバスやOSや、動作させるアプリケーションの要求性能に応じて、「ハードアレイ」方式のディスクアレイと、「ハイブリッドアレイ」方式のディスクアレイを、選択的に切り換えることができるディスクアレイコントローラを実現できるものである。

【0266】次に、図25～図26を用いて、本発明の第4の実施形態によるディスクアレイ制御システムについて説明する。従来のホスト計算機の内部バスに直結するタイプのディスクアレイ装置は、複数のホスト計算機により共用できないものであったが、本実施形態においては、これを可能とするものである。

【0267】最初に、図25を用いて、本発明の第4の実施形態によるマルチアレイ式のハイブリッドアレイ構成のディスクアレイ制御システムの全体構成について説明する。図25は、本発明の第4の実施形態によるディスクアレイ制御システムのハードウェアの全体構成のブロック図である。なお、図1及び図20と同一符号は、同一部分を示しており、以下の説明においては、図1との相違点を中心に説明する。

【0268】図25に示すように、本実施形態においては、2台のホスト計算機1a、1bを備えている。ホスト計算機1aは、図1に示したホスト計算機1と同一である。また、ホスト計算機1bは、図20に示したホスト計算機1と同一である。ホスト計算機1aは、ディスクアレイコントローラ3をPCI/F32で接続しており、ディスクアレイ装置2を内蔵している内蔵型の構成となっている。また、ホスト計算機1bは、ディスクアレイコントローラ3とSCSI33で接続している外付け型の構成となっている。

【0269】ディスクアレイコントローラ3のホストI/F選択手段37は、IF\_Mode[1:0]信号を送出する。ここで、図26を用いて、ホストI/F選択手段37が送出するIF\_Mode[1:0]信号について説明する。図26は、本発明の第4の実施形態によ

(25)

特開平10-105347

るディスクアレイ制御システムにおけるディスクアレイコントローラのホストI/F選択手段が送出するIF\_Mode[1:0]信号の論理図である。

【0270】ホストI/F選択手段37は、図26に示すようにIF\_Mode[1:0]信号をIF\_Mode0とIF\_Mode1の2本出力する。そして、IF\_Mode[1:0]信号=(0,1)の状態では、ホストI/Fとして、PCII/F32が選択され、IF\_Mode[1:0]信号=(1,0)の状態では、ホストI/Fとして、SCSI33が選択され、IF\_Mode[1:0]信号=(1,1)の状態では、ホストI/Fとして、PCII/F32とSCSI33との両方が選択されるクラスタモードとなるものである。即ち、ディスクアレイコントローラ3が備えている2つのホストI/Fの何れか一方、若しくは両方を使用可能にする。何れか一方を使用する際には、信号線が2本あることをのぞき、図1若しくは図20に示した実施形態と同様である。

【0271】また、両方使用する際には、ホストI/Fモード制御部305は、IF\_Mode[1:0]=(1,1)の状態では信号を受信すると、初期化部3035に通知し、初期化部3035はいずれのホストI/Fからのコントローラコマンドも処理できるようにPCII/F制御部301、およびSCSI制御部302を設定する。

【0272】また、ディスクアレイモード選択手段44により決定されたディスクアレイモードで動作しながら、両方のホストI/Fで受信したコントローラコマンドもしくは論理コマンドを処理するように、下位ディスクアレイ制御部303と第2上位ディスクアレイ制御部304を設定する。

【0273】この際、ディスクアレイモードに応じ、ホスト計算機は上位ディスクアレイ制御ドライバもしくはディスクドライバのいずれか一方を備える必要がある。

【0274】また、ディスクアレイモード選択手段44の出力信号DiskArray\_Mode信号も2本に拡張することで、各ホストI/F毎にディスクアレイモードを設定することも可能である。すなわち、例えば、ホスト計算機1aは、「ハイブリッドアレイモード」で動作し、ホスト計算機1bは、「ハードアレイモード」で動作することが可能である。

【0275】この際、PCII/F制御部301は、直接下位ディスクアレイ制御部303に受信したコントローラコマンドを引き渡し、また、SCSI制御部302は、第2上位ディスクアレイ制御部304に受信した論理コマンドを引き渡し、第2上位ディスクアレイ制御部304が、下位ディスクアレイ制御部303にコントローラコマンドを引き渡す。

【0276】以上のように、一台のディスクアレイ装置を、1台のホスト計算機1aに内蔵した状態でもう1台

のホスト計算機1bを同一ディスクアレイ装置に接続することができるので、ディスクアレイを2台のホスト計算機で共用できるようになる。

【0277】また、このように1台のディスクアレイ装置を2台のホスト計算機で共用できるので、2台のホスト計算機1a、1bの何れか一方を通常は動作していないスタンバイ機として運用し、普段動作する現用機が故障等によりダウンしたさいにスタンバイ機を動作させる、スタンバイ構成や、または両方のホスト計算機を常に動作させるクラスタ構成に対応するディスクアレイを構成できるものである。

【0278】また、上述した例では、ホスト計算機が2台の場合で説明したが、SCSIにさらに多くのホスト計算機を接続することで、n台のホスト計算機で唯一のディスクアレイを共用することができる。

【0279】以上のように、本実施形態によれば、「内蔵型ディスクアレイ」を実現した時にも、クラスタ構成や、スタンバイ構成等の複数台のホスト計算機で1台のディスクアレイを共用できるディスクアレイを実現できるようになる。

【0280】次に、図24及び図25を用いて、本発明の第5の実施形態によるディスクアレイ制御システムについて説明する。図25に示したディスクアレイコントローラ3を内蔵したホスト計算機1aの全体を一つのディスクアレイ装置とみなした例について説明する。この場合、ホスト計算機1aは、ホスト計算機としての機能は有しておらず、ユーザのアプリケーションプログラムや、ネットワークや、グラフィック処理等は行わないものである。

【0281】ここで、図25に示したディスクアレイコントローラ3のディスクアレイモード選択手段44を拡張し、図24に示すように、DiskArray\_ModeをDiskArray\_Mode[1:0]に2ビットに拡張することにより、「ハイブリッドアレイモード」及び「ハードアレイモード」に加えて、「超ディスクアレイモード」を設ける。この場合、SCSI33のみがホストI/Fとなる。

【0282】ディスクアレイモード選択手段44が送出するDiskArray\_Mode[1:0]=(1,1)の状態では、ディスクアレイモードとして、「超ディスクアレイモード」に初期設定する。

【0283】図25に示したホスト計算機1bのCPU10は、ディスクアレイコントローラ3に論理コマンドを発行する。図23において説明したように、SCSI制御部302は、論理コマンドを受信し、メモリ42に格納する。MPU30のPCII/F制御部301は、PCII/Fコントローラ321のDMA制御部を起動し、ホスト計算機1aの主記憶メモリ11にDMA転送する。

【0284】転送が終了すると、PCII/F制御部

(26)

特開平10-105347

はDoorBellレジスタにステータスを設定し、CPU10に割り込みを発行する。CPU10のPCI I/Fドライバは割り込みを受信すると、DoorBellレジスタのステータスをリードし、論理コマンドが転送されたことを認識する。上位ディスクアレイ制御ドライバ103は、論理コマンドを取り出し、以下、図1に示した実施形態と同様に動作する。

【0285】論理コマンドの実行が終了すると、上位ディスクアレイ制御ドライバ103はPCI I/Fドライバ321を制御し、論理コマンド終了ステータスを、図1に示した実施形態のコマンドリンク発行パケットの発行同様、ディスクアレイコントローラ3のメモリ42に転送する。MPU30のSCSI制御部302は、図23に示した実施形態と同様に、ホスト計算機1bに論理コマンドの終了を報告し、処理を終了する。

【0286】以上のように、ホスト計算機1aのCPU10は、上位ディスクアレイ制御ドライバを動作させることに専念できるので、さらに高性能化を図ることができる。また、より高度なディスクアレイ制御アルゴリズムを搭載し、処理をすることができるので一層の高性能化を図ることができる。

【0287】本実施形態によれば、ホスト計算機にディスクアレイを内蔵したときに、そのホスト計算機自体を一つのディスクアレイ装置として構成し、さらに高性能なディスクアレイを実現することができる。

【0288】

【発明の効果】本発明によれば、ディスクアレイコントローラに低価格なMPUを使用したディスクアレイ制御システムにおいて、MPUの能力で性能が制限されことなく高性能なディスクアレイ制御が実現できる。

【0289】また、本発明によれば、ディスクアレイコントローラに低価格なMPUを使用したディスクアレイ制御システムにおいて、MPUの能力で性能が制限されことなく高性能なディスクアレイ制御が実現できる。

【0290】さらに、本発明によれば、ホスト計算機のホストバスやOSや、動作させるアプリケーションの要求性能に応じて、「ハードアレイ」方式のディスクアレイと、「ハイブリッドアレイ」方式のディスクアレイを、選択的に切り換えることができるディスクアレイコントローラを実現できるものである。

【0291】また、さらに、本発明によれば、複数台のホスト計算機で1台の「内蔵型ディスクアレイ」を共用できるディスクアレイを実現でき、クラス構成やスタンバイ型構成を実現することができるものである。

【0292】さらに、本発明によれば、ホスト計算機にディスクアレイを内蔵したときに、そのホスト計算機自体を一つのディスクアレイ装置の一部として構成できるので、ホスト計算機の高性能なCPUを利用した高性能なディスクアレイを実現できるものである。

【図面の簡単な説明】

【図1】本発明の一実施形態によるディスクアレイ制御システムのハードウェアの全体構成のブロック図である。

【図2】本発明の一実施形態によるディスクアレイ制御システムのソフトウェアの全体構成のブロック図である。

【図3】本発明の一実施形態によるディスクアレイ制御システムのホストI/Fモード選択手段の構成を示す回路図である。

【図4】本発明の一実施形態によるディスクアレイ制御システムのホストI/Fモード選択手段の論理図である。

【図5】本発明の一実施形態によるディスクアレイ制御システムにおけるディスクアレイコントローラ検出処理を説明するフローチャートである。

【図6】本発明の一実施形態によるディスクアレイ制御システムにおけるディスクアレイのリード処理の説明図である。

【図7】本発明の一実施形態によるディスクアレイ制御システムにおける上位ディスクアレイ制御ドライバのコマンドリンクの発行方法について説明するフローチャートである。

【図8】本発明の一実施形態によるディスクアレイ制御システムにおけるホスト計算機とディスクアレイコントローラ間のコマンドリンクの発行方法の説明図である。

【図9】本発明の一実施形態によるディスクアレイ制御システムにおけるMPUのコマンドリンクの割り込み処理について説明するフローチャートである。

【図10】本発明の一実施形態によるディスクアレイ制御システムにおけるMPUのPCI I/F制御部の中のDMA制御部の処理について説明するフローチャートである。

【図11】本発明の一実施形態によるディスクアレイ制御システムにおけるMPUの下位ディスクアレイ制御部の処理について説明するフローチャートである。

【図12】本発明の一実施形態によるディスクアレイ制御システムにおけるMPUの下位ディスクアレイ制御部の処理について説明するフローチャートである。

【図13】本発明の一実施形態によるディスクアレイ制御システムにおけるMPUの下位ディスクアレイ制御部のディスクコマンド制御部の処理について説明するフローチャートである。

【図14】本発明の一実施形態によるディスクアレイ制御システムにおけるMPUの下位ディスクアレイ制御部のディスクコマンド制御部の終了報告受信時の処理について説明するフローチャートである。

【図15】本発明の一実施形態によるディスクアレイ制御システムにおけるホスト計算機のPCI I/Fドライバにおける終了報告処理について説明するフローチャートである。

(27)

特開平10-105347

ートである。

【図16】本発明の一実施形態によるディスクアレイ制御システムにおけるディスクアレイのライト処理の説明図である。

【図17】本発明の一実施形態によるディスクアレイ制御システムにおけるディスクアレイの第1のデステージ処理の説明図である。

【図18】本発明の一実施形態によるディスクアレイ制御システムにおけるディスクアレイの第2のデステージ処理の説明図である。

【図19】本発明の一実施形態によるディスクアレイ制御システムにおけるMPUの下位ディスクアレイ制御部のパリティ生成コマンド制御部の処理について説明するフローチャートである。

【図20】本発明の第2の実施形態によるディスクアレイ制御システムのハードウェアの全体構成のブロック図である。

【図21】本発明の第2の実施形態によるディスクアレイ制御システムのハードウェアのブロック斜視図である。

【図22】本発明の第2の実施形態によるディスクアレイ制御システムのホストI/Fモード選択手段の構成を示す回路図である。

【図23】本発明の第3の実施形態によるディスクアレイ制御システムのソフトウェアの全体構成のブロック図である。

【図24】本発明の第3の実施形態によるディスクアレイ制御システムにおけるディスクアレイコントローラのディスクアレイモード選択手段が送出するDiskArray\_Mode信号の論理図である。

【図25】本発明の第4の実施形態によるディスクアレイ制御システムのハードウェアの全体構成のブロック図である。

【図26】本発明の第4の実施形態によるディスクアレイ制御システムにおけるディスクアレイコントローラのホストI/F選択手段が送出するIF\_Mode[1:0]信号の論理図である。

【符号の説明】

1, 1a, 1b...ホスト計算機

10...CPU

101...OS

102...ファイルシステム

103...上位ディスクアレイ制御ドライバ

1031...論理コマンドアドレス変換部

1032...キャッシュ管理部

1034...コマンドI/F制御部

1035...コントローラ検出部

104...PCI I/Fドライバ

105...SCSIドライバ

106...ディスクドライバ

11...主記憶メモリ

12...システム制御手段

13...PCIバス

14...PCIコネクタ

15...SCSIコントローラ

16...SCSIコネクタ

17...SCSIケーブル

2...ディスクアレイ装置

21, 21a, 21b, 21c, 21d, 21e, 21f...ディスク

22...電源

23...クロック発生器

24...コネクタ

25...SCSI選択手段

26...SCSIコネクタ

28...筐体異常検出手段

29...ハードアレイ選択手段

3...ディスクアレイコントローラ

30...MPU

301...PCI I/F制御部

302...SCSI制御部

303...下位ディスクアレイ制御部

304...第2上位ディスクアレイ制御部

31...PCIカードエッジ

32...PCI I/F

321...PCI I/Fコントローラ

3211...DMAC

33, 34a, 34b...SCSI

331...SCSIコントローラ

332...SCSIコネクタ

34a1, 34b1...SCSIコントローラ

35...パリティ生成回路

36...キャッシュメモリ

37, 37a...ホストI/Fモード選択手段

38...PCIバス

40a1, 40b1...SCSIコネクタ

42...メモリ

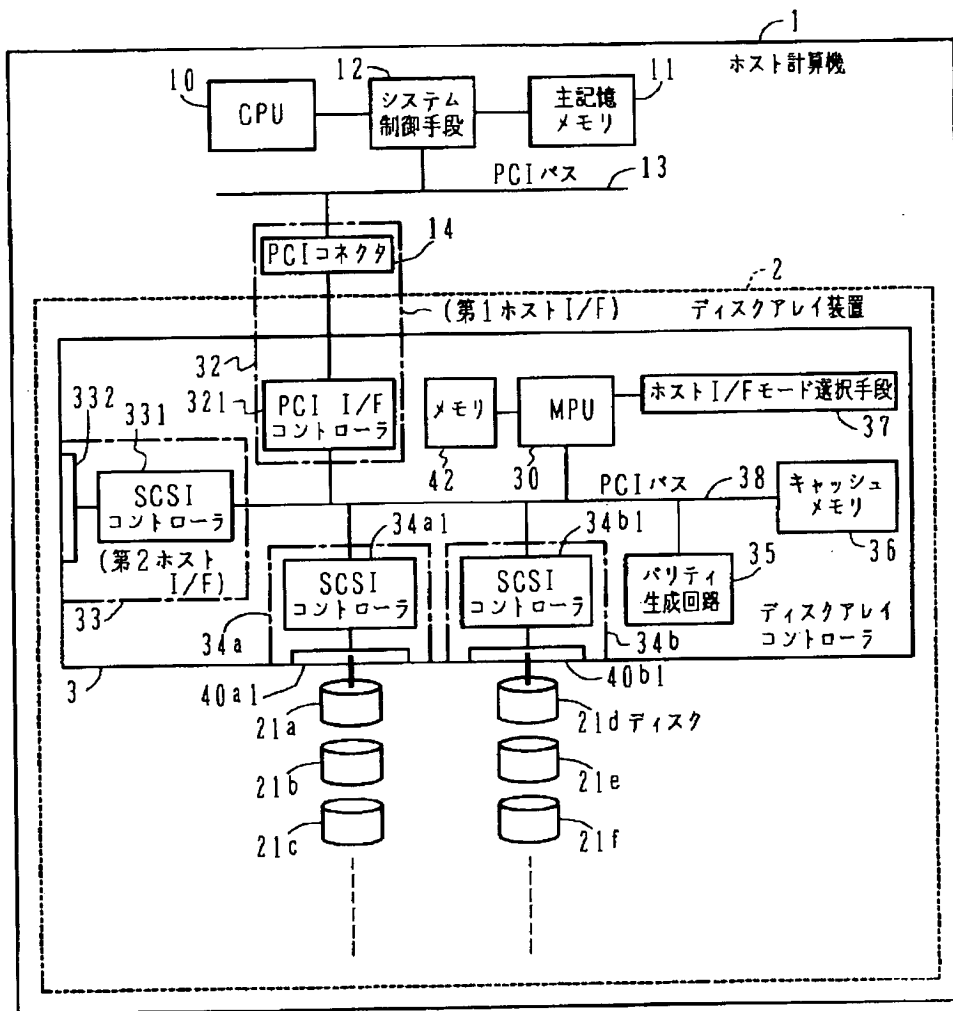
43...筐体監視手段

44...ディスクアレイモード選択手段

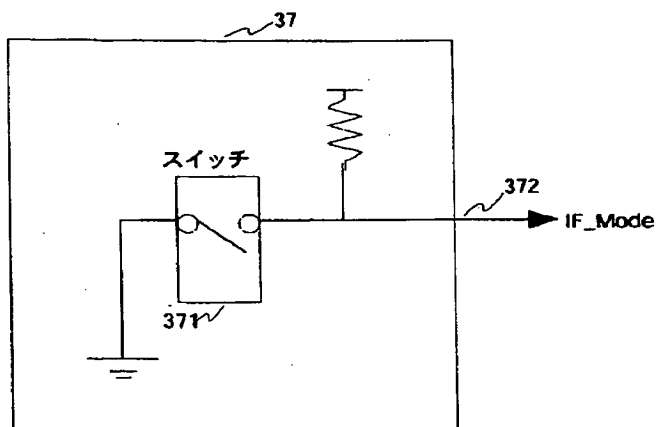
(28)

特開平10-105347

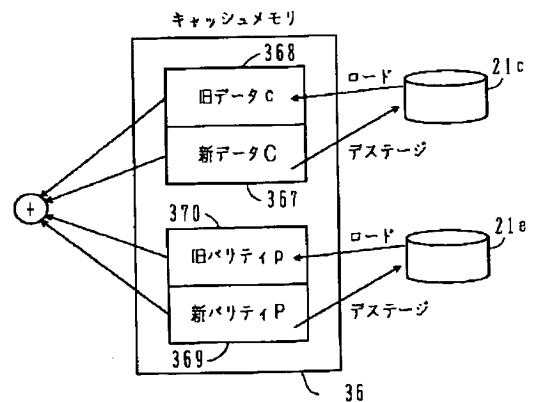
【図1】



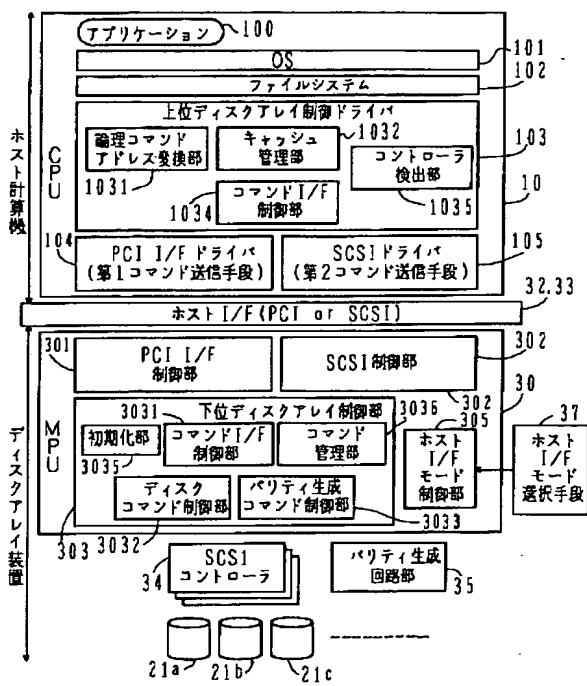
【図3】



【図17】



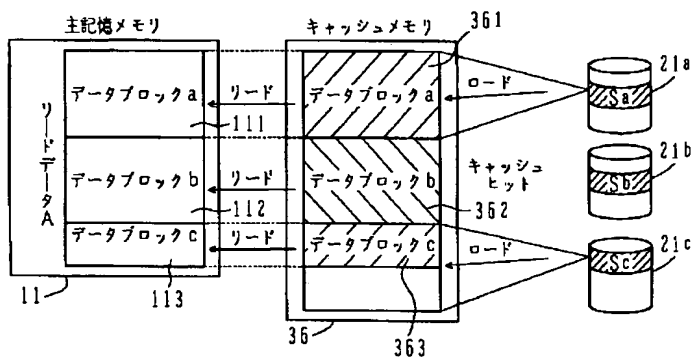
【図2】



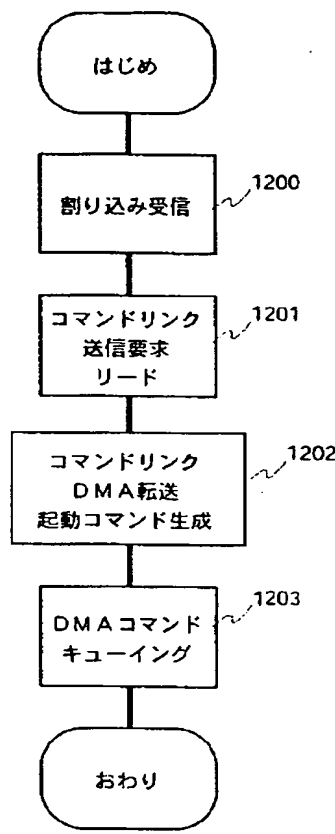
【図4】

IF_Mode	1	0
ホストI/F	PCI I/F (第一ホストI/F)	SCSI (第二ホストI/F)

【図6】



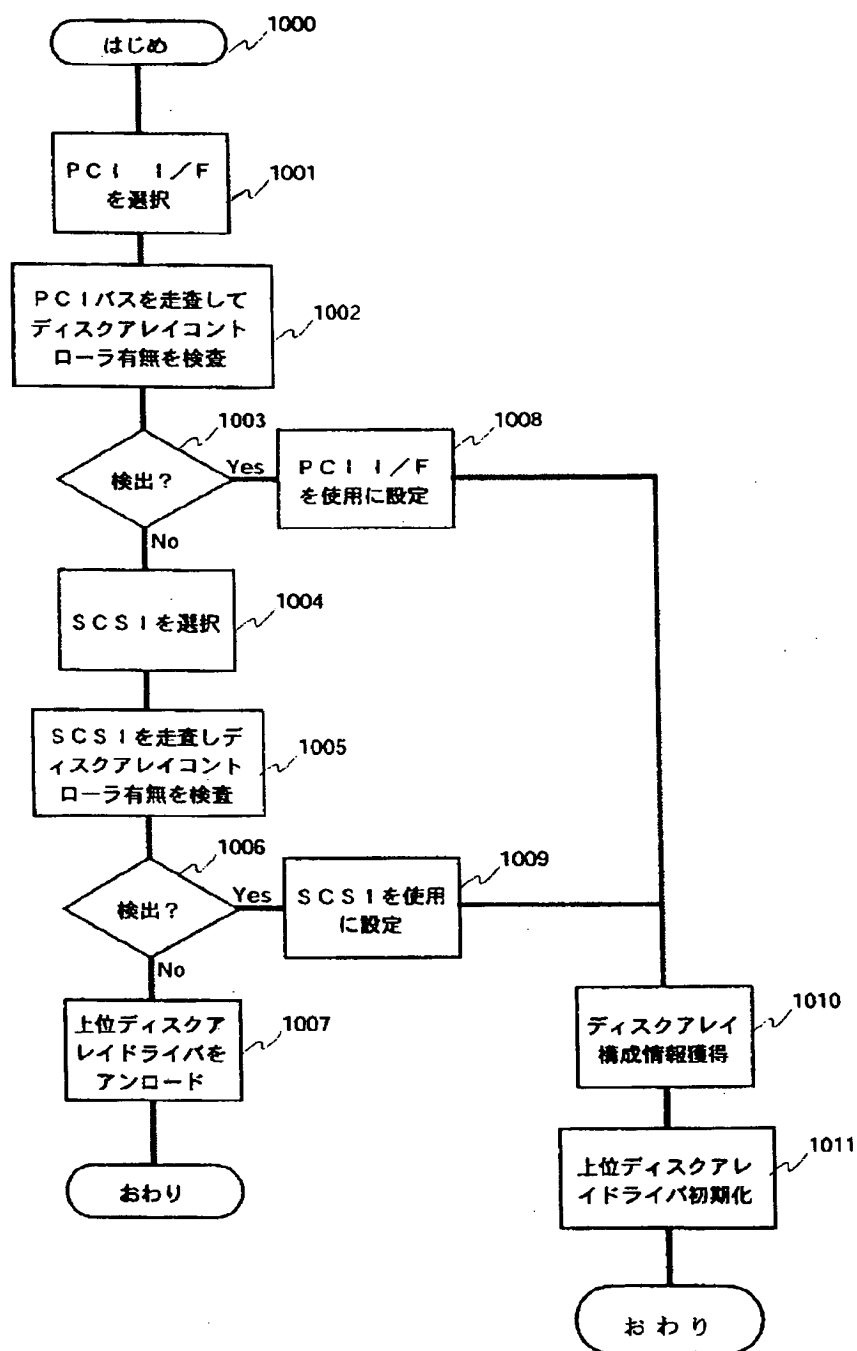
【図9】



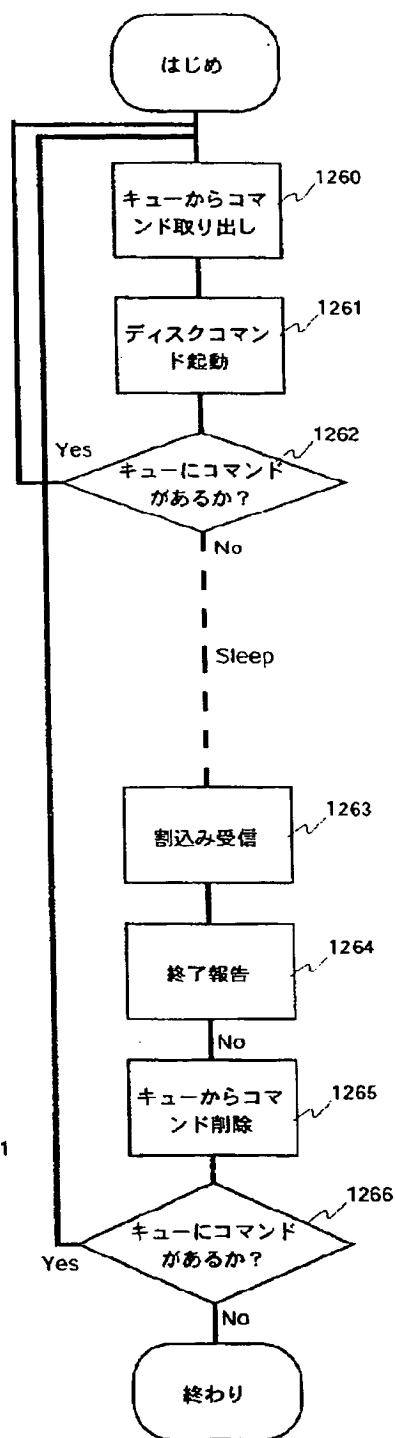
(30)

特開平10-105347

【図5】



【図13】

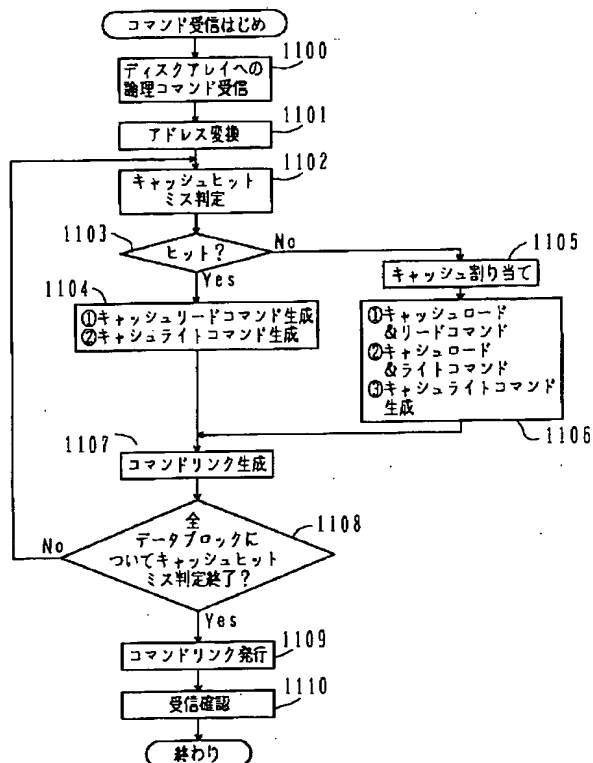




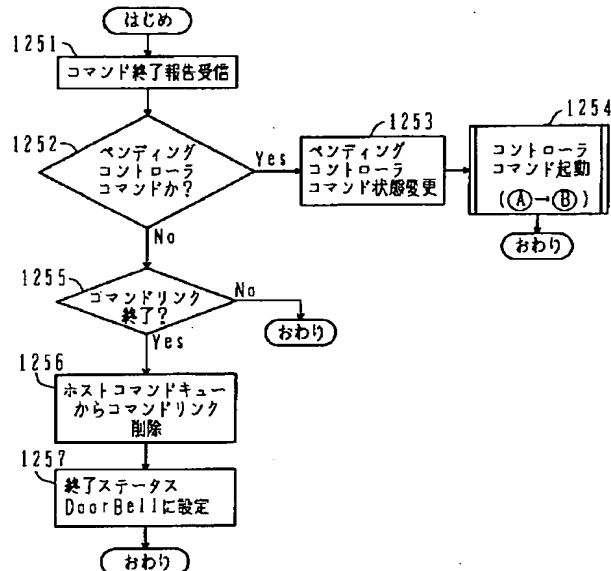
(31)

特開平10-105347

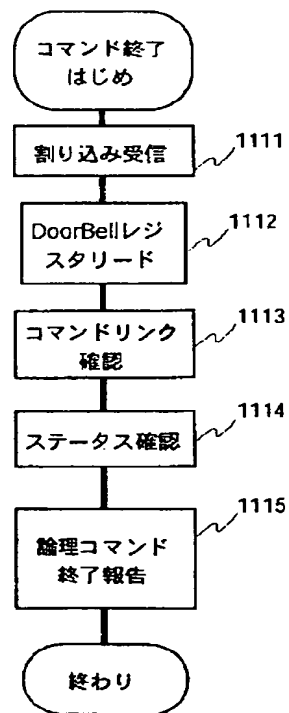
【図7】



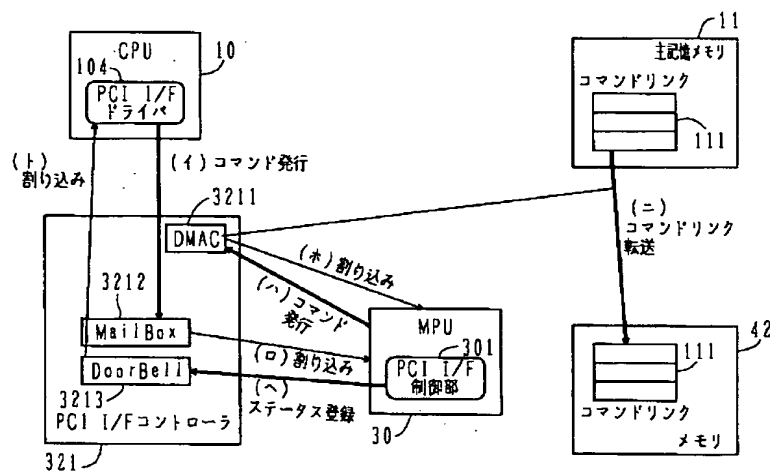
【図14】



【図15】



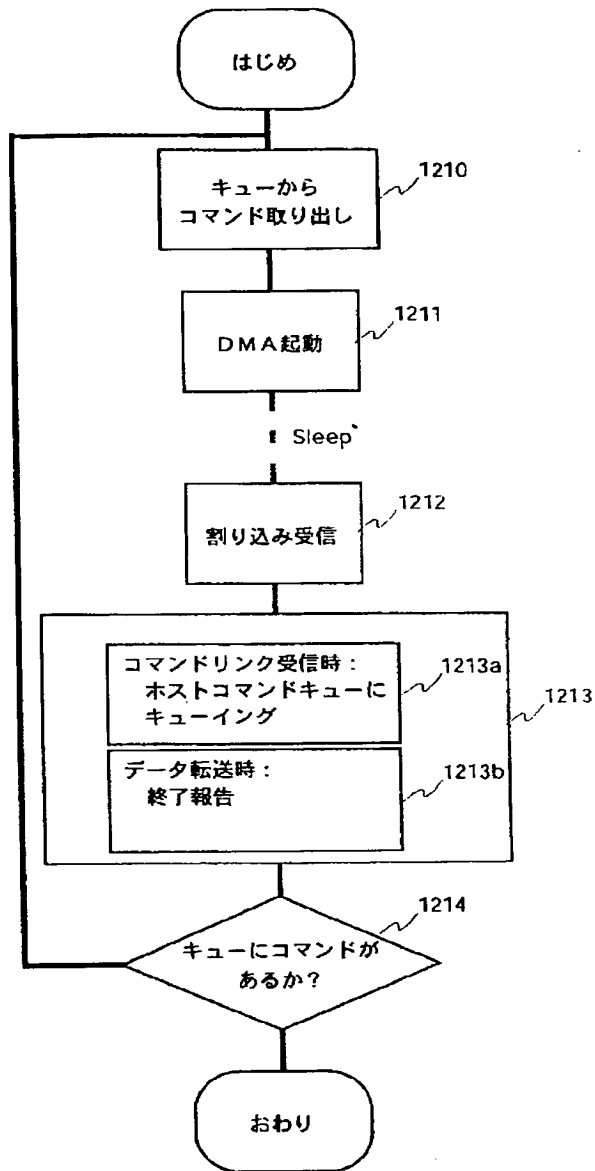
【図8】



(32)

特開平10-105347

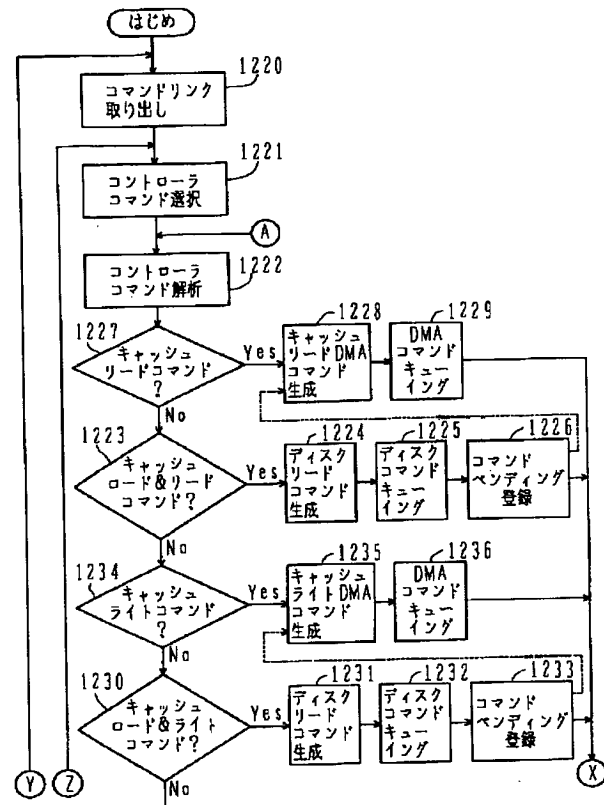
【図10】



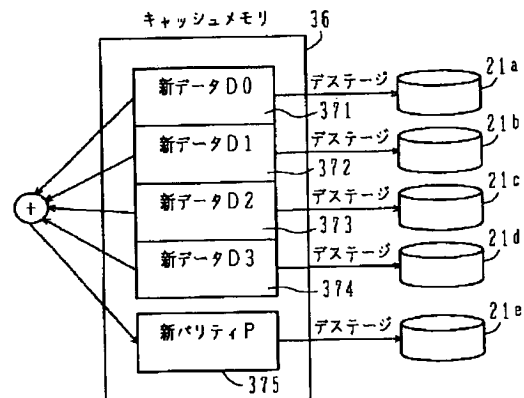
【図24】

DiskArray_Mode	1	0	DiskArray_Mode[1:0]
ディスクアレイ モード	ハイブリッド アレイモード	ハードアレイ モード	超ディスクアレイ モード

【図11】



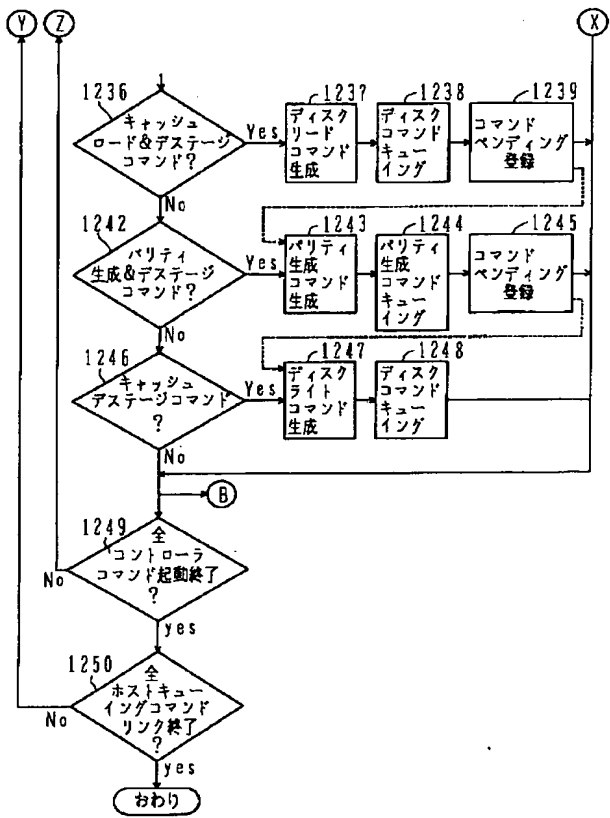
【図18】



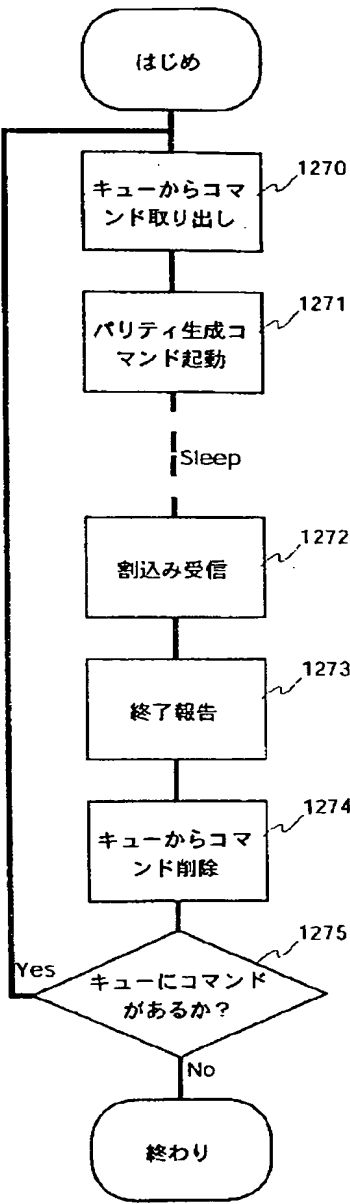
(33)

特開平10-105347

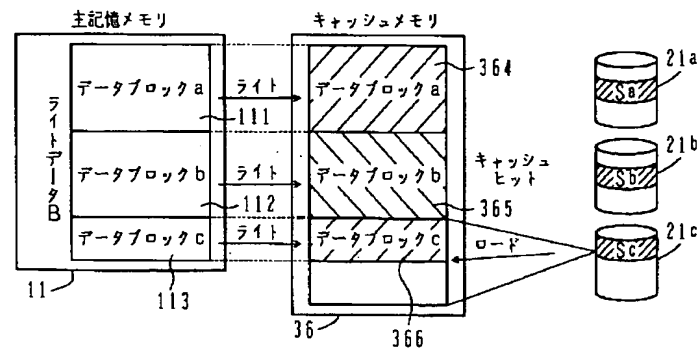
【図12】



【図19】



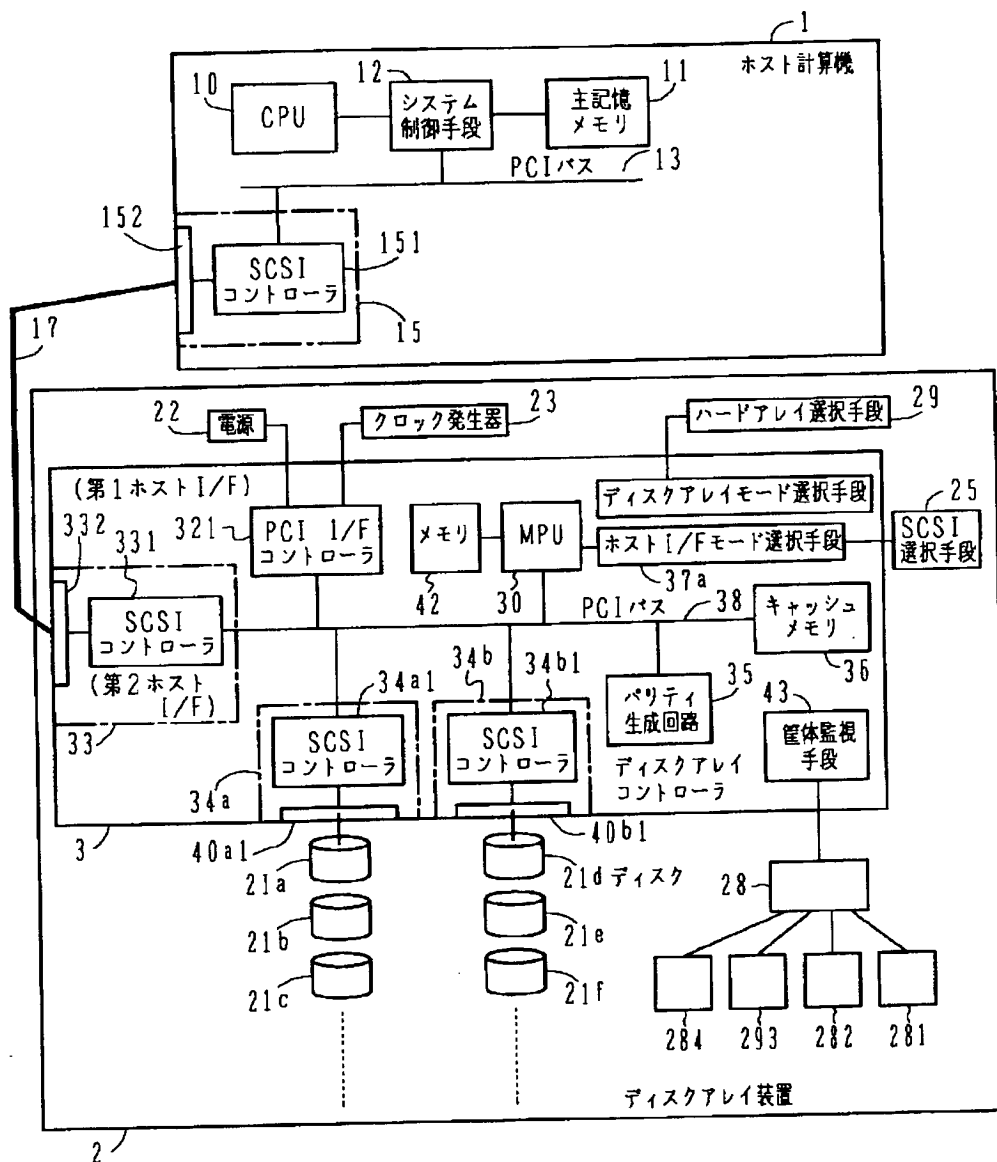
【図16】



(34)

特開平10-105347

【图20】



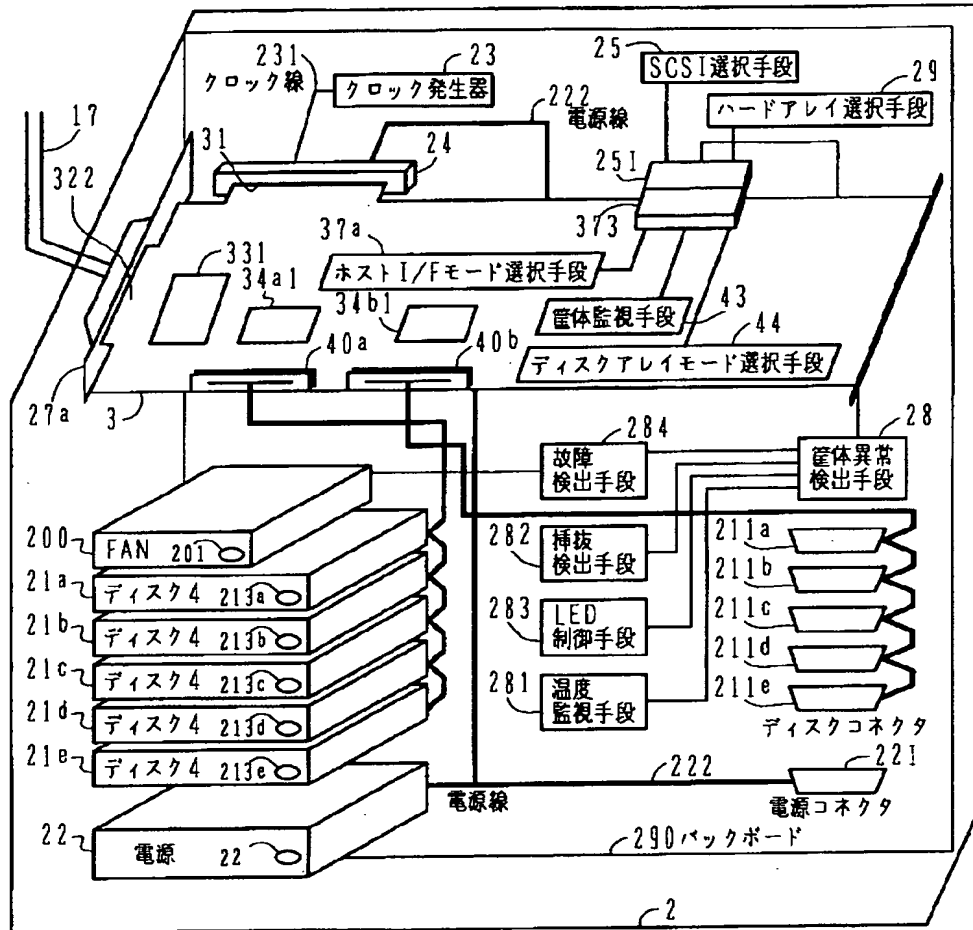
【图26】

IF_Mode[1:0]	1 1	1 0	0 1	0 0
ホスト I/F モード	両方使用 (クラスタモード)	SCSI 使用	PCI I/F 使用	未使用

(35)

特開平10-105347

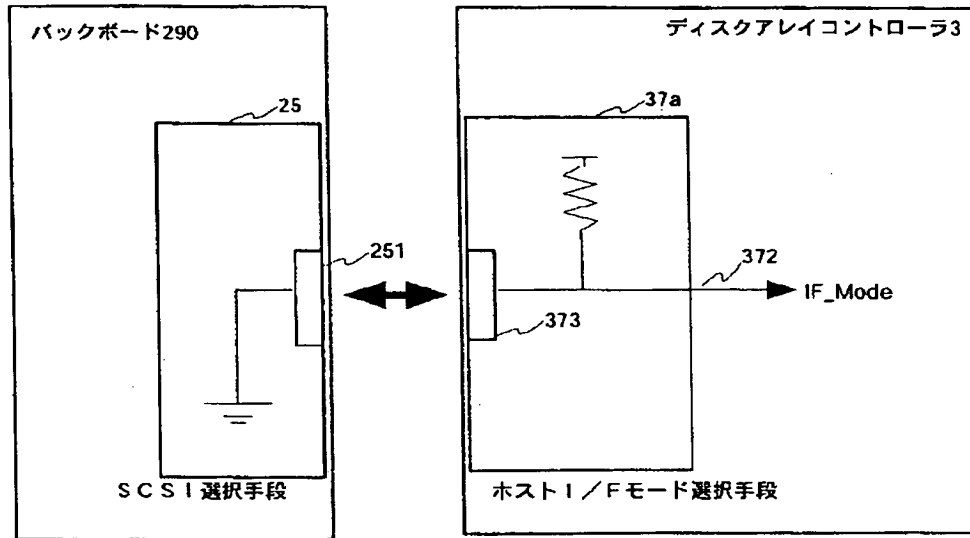
【図21】



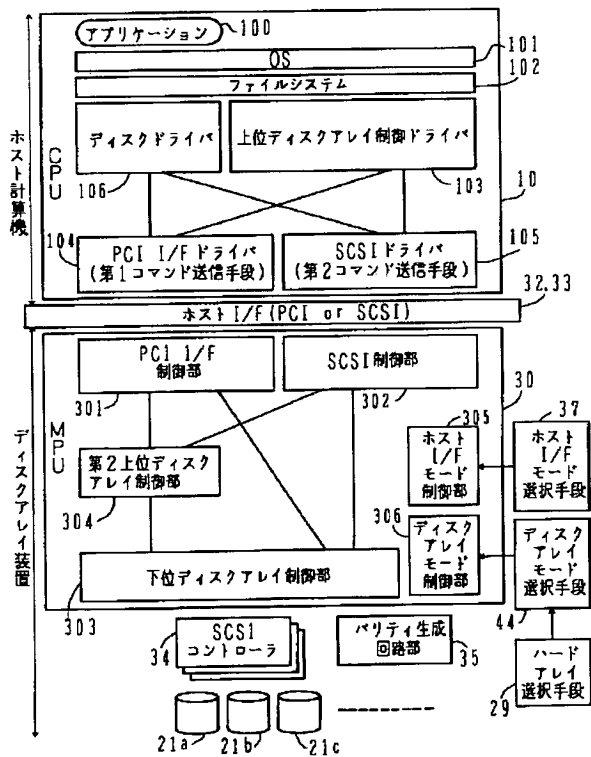
(36)

特開平10-105347

【図22】



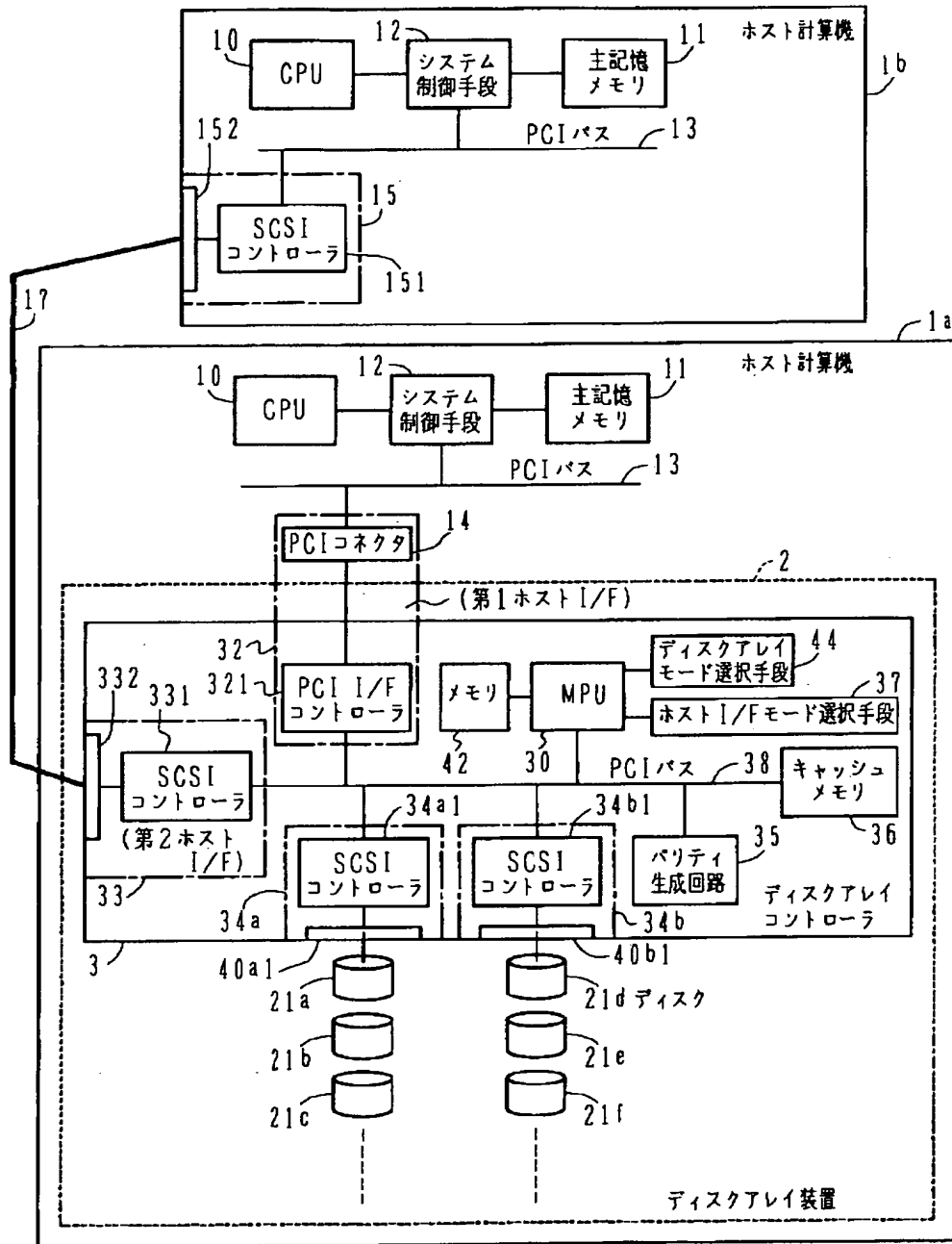
【図23】



(37)

特開平10-105347

【図25】



フロントページの続き

(72)発明者 荒川 敬史  
 神奈川県川崎市麻生区王禅寺1099番地 株  
 式会社日立製作所システム開発研究所内

(72)発明者 八木沢 育哉  
 神奈川県川崎市麻生区王禅寺1099番地 株  
 式会社日立製作所システム開発研究所内

(38)

特開平10-105347

(72)発明者 高野 雅弘  
神奈川県小田原市国府津2880番地 株式会  
社日立製作所ストレージシステム事業部内